Dynamic Path Analysis and Model based clustering of microarray data

by Matthias Kormaksson

DYNAMIC PATH ANALYSIS AND MODEL BASED CLUSTERING OF MICROARRAY DATA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Matthías Kormáksson

August 2010

DYNAMIC PATH ANALYSIS AND MODEL BASED CLUSTERING OF MICROARRAY DATA

Matthías Kormáksson, Ph.D.

Cornell University 2010

Part I

We consider a situation where we observe continuous and binary data for different subjects at discrete time points. At each time point the binary responses are modeled with probit equations and the continuous responses with linear regression equations. The model construction results in a Gaussian system of equations with directed acyclic graph structure, where the variables, as well as the parameters, are time-dependent. The functional parameters are further modeled with a mixed model representation of splines and estimation is carried out with a Bayesian analysis. We establish a connection with dynamic graphical models and through a simple Gibbs sampler we obtain posterior estimates of direct, indirect and total effects of our model. These estimates allow us to describe how the effects of fixed covariates are working partly directly and partly indirectly through endogenous time-dependent covariates. We show how our methodology can be applied in certain situations arising in Survival analysis and we illustrate our methods on a simple data set.

Part II

A recent study on a cohort of 344 well-characterized patients with acute myeloid leukemia suggests that subjects can be segregated into distinct groups using unsupervised clustering based on their DNA methylation profiles. We suggest a model based approach, where we introduce latent cluster specific methylation indicators on each gene. These indicators along with some standard assumptions impose a specific mixture distribution on each cluster and the parameters of the induced likelihood are estimated using the EM algorithm. We also introduce latent gene importance indicators, which provides us with information about which genes discriminate between patients. By calculating posterior expectations of the above indicators we can predict genomewide methylation patterns across different subtypes of AML, which facilitates AML classifications of new patients based on their methylation profiles. The methods we develop extend naturally to other data types of similar nature such as expression data. This leads to a joint analysis over multiple data platforms, resulting in a higher discriminating power.

BIOGRAPHICAL SKETCH

Matthías Kormáksson was born May 6th 1981 in Saudárkrókur, a small town in the northern most part of Iceland. At the age of four he moved to Reykjavik and has since lived in the surroundings of the capital. In the years 1997-2001 he was a student at the secondary school Menntaskólinn í Reykjavík, where he majored in Math and Physics. In the Fall of 2001, he enrolled in the Math department at the University of Iceland, where he graduated in the spring of 2004 with a B.Sc. degree. He was accepted into the Ph.D. program at the Department of Statistical Science, Cornell University in the Fall of 2004, and finished his degree in August 2010.

Þessi ritgerð er tileinkuð foreldrum mínum.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my adviser, Professor James Booth, for his countless suggestions and insightful hints while conducting my research and writing this thesis. His intuition on the material has been inspirational and has greatly helped me improve my skills as a Statistician.

I would also like to thank the other two members of my special committee: Professor Robert Strawderman, for the time and effort he put into valuable comments for improving my thesis, and Professor Marty Wells whose comments and suggestions during my research have been very useful.

Finally, I would like to thank my dear friends in the Linux lab and at the Chapter House, for sharing different perspectives on life both inside and outside of Cornell.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

Dynamic path analysis

CHAPTER 1

INTRODUCTION

Graphical models are widely used in statistics and have applications in several fields, such as genetics, econometrics and social sciences. The mathematical theory of these models is now well developed and great overviews of theory and applications can be found in for example Lauritzen (1996) and Edwards (2000). One of the appealing features of graphical models is the easily communicated graphical representation of the models. Furthermore the notion of direct, indirect and total effects give us a clearer picture of interrelations between variables. One factor is mostly ignored in the classical graphical models literature and that is the role of time. In many situations one could imagine a system of equations that evolves over time with model effects changing dynamically. In this part of the thesis we develop a method for estimating time changing parameters in such dynamic graphical models. We focus on models with acyclic directed graph structure. This problem is inspired by the work of Fosen et al. (2006), which focused on dynamic systems arising in survival analysis. In their paper the main outcome of interest was a counting process and its compensator was modeled as a linear function of a set of covariates, using Aalen's additive hazard model (Aalen (1980)). The covariates themselves were then modeled through a system of linear equations and estimation carried out separately for each equation and each time point. The problem with estimating the system of equations separately at each time point is that it does not give a smooth estimate of the time changing parameters.

In chapter 2 we propose a model that can be applied to several endogenous continuous variables and a single endogenous binary variable, observed for different subjects at discrete time points. With a proper construction the methodology can be used for analyzing certain types of survival data. At each time point the continuous variables are

modeled with linear models but the binary variable is modeled with a probit equation. We introduce a latent Gaussian variable, corresponding to the binary variable, that is > 0 whenever the binary variable takes on the value 1, and ≤ 0 when it takes on the value 0. This allows us to transform the system of equations, involving the linear models and the probit equation, into a Gaussian system of equations. The idea of a latent continuous process underlying a binary response can for example be found in Gueorguieva and Agresti (2001). We further impose a functional structure on the time varying parameters, with the use of splines. This results in a simple Gaussian mixed model whose parameters we estimate with a Gibbs sampler. At each step of the Gibbs sampler we impute the missing latent variable by sampling from a truncated normal random variable. The estimation procedure is detailed in chapter 3 but we consider the case of independent errors in the mixed model as well as AR(1) dependence structure. We end chapter 3 by showing how one can construct simultaneous credible bands around the estimated functional parameters of the model. In chapter 4 we do a simple analysis of the head and neck cancer study conducted by The Northern California Oncology Group. The data is described in detail and analyzed by Efron (1988) but we show how the model proposed there is a special case of the model presented here.

CHAPTER 2

DYNAMIC PATH ANALYSIS WITH CONTINUOUS AND BINARY DATA

Assume we observe continuous and binary data over time. We have G longitudinal endogenous variables $Y_{ik}(t)$, $k = 1, \ldots, G$ and a single longitudinal binary variable $s_i(t)$, $i = 1, \ldots, n$ where n denotes the number of units under observation. The processes are observed at the discrete time points t_{ij}, $j = 1, \ldots, m_i$ not necessarily equally spaced. The timings of observations are assumed independent of all observation processes. In this part of the thesis we will at each time point consider a system of linear equations where the first G equations involve $Y_{ik}(t)$, $k = 1, \ldots, G$ as the responses and the last equation represents a probit model for the binary response $s_i(t)$. We will furthermore assume that our system of equations has a directed acyclic graph structure, which we will explain below. In the first section of this chapter we will introduce the general model. In the second section we will talk about the relationship between our model and Dynamic Graphical models as well as the notion of direct and indirect effects and in the final section we will show how the model can be applied to survival data with a proper construction.

2.1 The Model

Assume we observe, additionally to the endogenous continuous and bivariate variables, the exogenous covariates $\mathbf{X}_i(t) = (X_{i1}(t), \ldots, X_{iK}(t))$ for each unit $i = 1, \ldots, n$. Define $h_i(t_{ij}) := \mathrm{P}\{s_i(t_{ij}) = 1 | Y_{i1}(t_{ij}), \ldots, Y_{iG}(t_{ij}), \mathbf{X}_i(t_{ij})\}$ and consider the following

model

$$\begin{aligned} Y_{i1}(t) &= \mathbf{X}_i^{(1)}(t)\boldsymbol{\delta}_1(t) + \varepsilon_1(t) \\ Y_{i2}(t) &= \mathbf{Y}_i^{(2)}(t)\boldsymbol{\gamma}_2(t) + \mathbf{X}_i^{(2)}(t)\boldsymbol{\delta}_2(t) + \varepsilon_2(t) \\ &\vdots \\ Y_{iG}(t) &= \mathbf{Y}_i^{(G)}(t)\boldsymbol{\gamma}_G(t) + \mathbf{X}_i^{(G)}(t)\boldsymbol{\delta}_G(t) + \varepsilon_G(t) \\ \Phi^{-1}\left(h_i(t)\right) &= \mathbf{Y}_i^{(G+1)}(t)\boldsymbol{\gamma}_{G+1}(t) + \mathbf{X}_i^{(G+1)}(t)\boldsymbol{\delta}_{G+1}(t) \end{aligned} \tag{2.1}$$

where the row vector $\mathbf{X}_i^{(j)}(t)$ denotes the subvector of $\mathbf{X}_i(t) = (X_{i1}(t), \ldots, X_{iK}(t))$ that consists of the exogenous covariates occurring in the j-th equation. Similarly the row vector $\mathbf{Y}_i^{(j)}(t)$ denotes the subvector of $(Y_{i1}(t), \ldots, Y_{i,j-1}(t))$ that consists of the endogenous covariates occurring in the j-th equation. Notice the difference between the two definitions. We are assuming that our system of equations has a directed acyclic graph structure, which means that the j-th equation can only contain endogenous variables from the subvector $(Y_{i1}(t), \ldots, Y_{i,j-1}(t))$ but not the whole vector of all the endogenous variables, $(Y_{i1}(t), \ldots, Y_{iG}(t))$. The notion of a directed acyclic graph (DAG) is addressed fully in section 3 of this chapter. The vectors $\boldsymbol{\gamma}_j(t) = (\gamma_{j1}(t), \ldots, \gamma_{jg_j}(t))'$ and $\boldsymbol{\delta}_j(t) = (\delta_{j1}(t), \ldots, \delta_{jd_j}(t))'$ are the corresponding parameter vectors, where g_j and d_j denote the number of endogenous and exogenous variables respectively in the j-th equation. The function Φ denotes the standard normal cumulative density function.

Now for each t let us define a latent normal random variable $Y_{i,G+1}(t)$, such that the observed binary response is the indicator $s_i(t) = I(Y_{i,G+1}(t) > 0)$. We consider the

following Gaussian system of equations:

$$\begin{aligned} Y_{i1}(t) &= \mathbf{X}_i^{(1)}(t)\boldsymbol{\delta}_1(t) + \varepsilon_{i1}(t) \\ Y_{i2}(t) &= \mathbf{Y}_i^{(2)}(t)\boldsymbol{\gamma}_2(t) + \mathbf{X}_i^{(2)}(t)\boldsymbol{\delta}_2(t) + \varepsilon_{i2}(t) \\ &\vdots \\ Y_{iG}(t) &= \mathbf{Y}_i^{(G)}(t)\boldsymbol{\gamma}_G(t) + \mathbf{X}_i^{(G)}(t)\boldsymbol{\delta}_G(t) + \varepsilon_{iG}(t) \\ Y_{i,G+1}(t) &= \mathbf{Y}_i^{(G+1)}(t)\boldsymbol{\gamma}_{G+1}(t) + \mathbf{X}_i^{(G+1)}(t)\boldsymbol{\delta}_{G+1}(t) + \varepsilon_{i,G+1}(t) \end{aligned} \tag{2.2}$$

It is not hard to see that with $\varepsilon_{i,G+1}(t) \sim \mathrm{N}(0,1)$ this model translates directly into the correlated probit model in (2.1). Note that the $G+1$ error variance in the above model cannot be identified jointly with the parameters in the $G+1$ equation and is set equal to 1 for convenience. This makes sense as the only information about the parameters and variance is contained in $s_i(t) = \mathrm{I}(Y_{i,G+1}(t) > 0)$. Take the two sets of parameters $(\boldsymbol{\gamma}(t), \boldsymbol{\delta}(t), \sigma^2_{G+1})$ and $(\boldsymbol{\gamma}(t)' = \frac{\boldsymbol{\gamma}(t)}{\sigma_{G+1}}, \boldsymbol{\delta}(t)' = \frac{\boldsymbol{\delta}(t)}{\sigma_{G+1}}, 1)$ and note that both of these will result in the same probability value of $\mathrm{P}\{s_i(t) = 1 | Y_{i1}(t), \ldots, Y_{iG}(t), \mathbf{X}_i(t)\}$. The goal is now to estimate all the functional parameters $\gamma_{jk}(t)$ and $\delta_{jk}(t)$ in (2.2).

Let $\mathbf{Y}_i(t) = (Y_{i,1}(t), \ldots, Y_{i,G}(t), Y_{i,G+1}(t))'$ and $\boldsymbol{\varepsilon}_i(t) = (\varepsilon_{i1}(t), \ldots, \varepsilon_{i,G+1}(t))'$ be the response and error vector for subject, $i = 1, \ldots, n$, at time, t, where n is the total number of subjects. Let $\mathbf{S}_i^{(1)}(t) = \mathbf{X}_i^{(1)}(t)$ and $\mathbf{S}_i^{(j)}(t) = [\mathbf{Y}_i^{(j)}(t) \quad \mathbf{X}_i^{(j)}(t)]$ for $2 \leq j \leq G+1$. Then define $\mathbf{S}_i(t) = \text{blockdiag}(\mathbf{S}_i^{(j)}(t))$ and let $\boldsymbol{\eta}(t) = (\boldsymbol{\delta}_1(t)', \boldsymbol{\gamma}_2(t)', \boldsymbol{\delta}_2(t)', \ldots, \boldsymbol{\gamma}_{G+1}(t)', \boldsymbol{\delta}_{G+1}(t)')'$. We note that for a fixed t the vector $\boldsymbol{\eta}(t)$ is of length $g+d$ where $g = g_2 + \cdots + g_{G+1}$ and $d = d_1 + \cdots + d_{G+1}$. The matrix $\mathbf{S}_i(t)$ is a $(G+1) \times (g+d)$-matrix. With the above simplifications the model (2.2) can now be written in the following manner:

$$\mathbf{Y}_i(t) = \mathbf{S}_i(t)\boldsymbol{\eta}(t) + \boldsymbol{\varepsilon}_i(t) \tag{2.3}$$

which is simply a system of time-varying coefficients equations. For ease of notation we will refer to the components of $\boldsymbol{\eta}(t)$ as $\eta_1(t), \ldots, \eta_{g+d}(t)$.

Next we model the components of the time varying coefficients vector $\boldsymbol{\eta}(t)$. We propose a mixed model representation of a spline

$$\eta_k(t) = \mathbf{B}(t)\boldsymbol{\beta}_k + \mathbf{C}(t)\mathbf{u}_k, \quad k = 1, \ldots, g+d \tag{2.4}$$

where the $\mathbf{u}_k$'s are considered normally distributed and for now we can think of the $\boldsymbol{\beta}_k$'s as fixed. Let us assume for simplicity that the vectors $\mathbf{B}(t)$ and the $\mathbf{C}(t)$ are the same for all $g+d$ coefficients. Of course one could allow them to depend on k to allow for more flexibility in modeling the different coefficient curves. An example of how we could specify $\mathbf{B}(t)$ and $\mathbf{C}(t)$ is to let

$$\mathbf{B}(t) = (1, t, t^2, \ldots, t^p) \quad \text{and} \quad \mathbf{C}(t) = ((t-\tau_1)_+^p, \ldots, (t-\tau_\kappa)_+^p)$$

where $\tau_1, \ldots, \tau_\kappa$ are internal knots within the interval $(0, \tau)$. It follows from (2.4) that

$$\boldsymbol{\eta}(t) = (\mathbf{I}_{g+d} \otimes \mathbf{B}(t))\boldsymbol{\beta} + (\mathbf{I}_{g+d} \otimes \mathbf{C}(t))\mathbf{u}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_{g+d}')'$ and $\mathbf{u} = (\mathbf{u}_1', \ldots, \mathbf{u}_{g+d}')'$. The above mixed model representation of the functional parameters has been established in the context of smoothing. Ruppert et al. (2003) discuss in detail the estimation of the function f in the nonparametric regression model:

$$y_i = f(x_i) + \varepsilon_i, \quad 1 \le i \le n$$

and show that by representing f with the spline

$$f(x_i) = \beta_0 + \beta_1 x_i + \ldots \beta_p x_i^p + \sum_{k=1}^{K} u_k (x_i - \tau_k)_+^p$$

the estimation can be carried out by assuming the u_k's are random and fitting the mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

where $\mathbf{X}$ is the matrix whose ith row is polynomial basis $(1, x_i, \ldots, x_i^p)$ and $\mathbf{Z}$ is the matrix whose ith row is the truncated basis $((x_i - \kappa_1)_+^p, \ldots, (x_i - \kappa_K)_+^p)$. The mixed

model estimation is shown to be equivalent to the estimation of f by the use of penalized splines. The above technique fits quite naturally into our framework of estimating time-varying coefficients. By defining $\mathbf{W}_i(t) = \mathbf{S}_i(t)(\mathbf{I}_{g+d} \otimes \mathbf{B}(t))$ and $\mathbf{Z}_i(t) = \mathbf{S}_i(t)(\mathbf{I}_{g+d} \otimes \mathbf{C}(t))$ the model (2.3) simply becomes the mixed model

$$\mathbf{Y}_i(t) = \mathbf{W}_i(t)\boldsymbol{\beta} + \mathbf{Z}_i(t)\mathbf{u} + \boldsymbol{\varepsilon}_i(t) \tag{2.5}$$

It is important to note a subtle but crucial difference between this setup and the smoothing procedure of Ruppert et al. (2003). In our model the design matrices $\mathbf{W}_i(t)$ and $\mathbf{Z}_i(t)$ actually depend on the covariates $(\mathbf{Y}_i^{(j)}(t), \mathbf{X}_i^{(j)}(t))_j$. This requires special care when developing the fitting procedure in chapter 3.

Let us now stack the observations across time for individual i. Define $\mathbf{Y}_i = (\mathbf{Y}_i(t_{i1})', \ldots, \mathbf{Y}_i(t_{im_i})')'$, and $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_i(t_{i1})', \ldots, \boldsymbol{\varepsilon}_i(t_{im_i})')'$. Let $\mathbf{W}_i$ and $\mathbf{Z}_i$ be the matrices that are constructed by stacking the matrices $\mathbf{W}_i(t_{ij})$ and $\mathbf{Z}_i(t_{ij})$, $j = 1, \ldots, m_i$, on top of one another. Then from (2.5) we get the following mixed model

$$\mathbf{Y}_i = \mathbf{W}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u} + \boldsymbol{\varepsilon}_i \tag{2.6}$$

Finally define $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_n')'$ and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_n')'$ and let $\mathbf{W}$ and $\mathbf{Z}$ be the matrices that are obtained by stacking the matrices $\mathbf{W}_i$ and $\mathbf{Z}_i$, $i = 1, \ldots, n$, on top of one another. Then from (2.6) we obtain

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \tag{2.7}$$

which is our model to be fitted. Where appropriate we will also refer to the models (2.5) and (2.6). Recall that the components of $\mathbf{u}$ were considered normal. It is logical to assign a different variance component, $\sigma_{u\ell}^2$ to each component vector $\mathbf{u}_\ell$ of $\mathbf{u}$, $l = 1, \ldots, g + d$. But as is customary to ease computations we will assume the components of $\mathbf{u}$ are all independent. More precisely we assume

$$\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{D}_u \otimes \mathbf{I}_\kappa)$$

where we define $\mathbf{D}_u = \text{diag}_{1\leq\ell\leq g+d}(\sigma^2_{u\ell})$. Note that the total number of variance components $\sigma^2_{u\ell}$ is bounded above by $g + d \leq G(G+1)/2 + K(G+1)$.

As for the error structure we assume normality and homogeneity of variance across time. To be more precise we assume that each error vector $\boldsymbol{\varepsilon}_i(t)$ in (2.5) is normally distributed with a covariance matrix $\boldsymbol{\Sigma}$ independent of t. If we regard the right hand sides of (2.2) as conditional means plus error terms the directed acyclic graph structure imposes independence on the errors, see appendix A.2. This means that we can assume a diagonal covariance structure $\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{\varepsilon}_i(t_{ij})) = \text{diag}_{1\leq k\leq G+1}(\sigma^2_k)$, for all i, j, where σ^2_{G+1} is constrained to be equal to 1. We will consider two specific covariance structures for the individual error vectors, $\boldsymbol{\varepsilon}_i$, in (2.6), but will otherwise assume that $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ for the n individuals are independent. Firstly, we will consider the simple case of independence across time, where we let $\boldsymbol{\Omega}_i := \text{Cov}(\boldsymbol{\varepsilon}_i) = \mathbf{I}_{m_i} \otimes \boldsymbol{\Sigma}$ and by the independence between individuals $\boldsymbol{\Omega} := \text{Cov}(\boldsymbol{\varepsilon}) = \text{blockdiag}(\boldsymbol{\Omega}_i)$. Secondly we will assume that the errors are autocorrelated across time. The estimation procedures for both scenarios are considered in chapter 3.

2.2 Dynamic Graphical Models - Direct and Indirect Effects

A graph, $\mathcal{G} = (V, E)$, is a structure consisting of a vertex set V and an edge set E containing the edges between vertices. In graphical models the vertices represent random variables and the edges formulate the distributional relationship between the vertices according to some predetermined rules. For a thorough overview of graphical models,see Edwards (2000) or Lauritzen (1996). A directed graph is a graph $\mathcal{G} = (V, E)$, where $E \subset V \times V$ is now a set of ordered pairs of vertices. If $Y_1, Y_2 \in V$ are two vertices, then $(Y_1, Y_2) \in E$ is not the same as $(Y_2, Y_1) \in E$. We usually write $Y_1 \to Y_2$ instead of

$(Y_1, Y_2) \in E$ and say that Y_1 is a parent of Y_2 and Y_2 is a child of Y_1. We denote the set of parents of a vertex Y as pa(Y). A path in the directed graph is defined to be a sequence of vertices $\{Y_1, \ldots, Y_k\}$ such that $Y_j \to Y_{j+1}$ or $Y_{j+1} \to Y_j$ for all $j = 1, \ldots, k-1$. A directed path from Y_1 to Y_k is a path such that $Y_j \to Y_{j+1}$ for each $j = 1, \ldots, k-1$. We sometimes denote a directed path by $Y_1 \to Y_2 \to \cdots \to Y_k$. When $Y_1 = Y_k$ the directed path $Y_1 \to Y_2 \to \cdots \to Y_k$ is called a directed cycle. This brings us to the definition of the graphical structure considered in this part of the thesis.

Definition 2.2.1. *A directed acyclic graph (DAG) is a directed graph $\mathcal{G} = (V, E)$ with no directed cycles.*

In constructing a probabilistically meaningful DAG, we consider a set of random variables $Y_1, \ldots, Y_k$ and let them constitute the vertex set V. The joint density of $Y_1, \ldots, Y_k$ can be factored into

$$f(Y_1)f(Y_2|Y_1) \cdot \cdots \cdot f(Y_k|Y_1, \ldots, Y_{k-1})$$

and for $i < j$ we draw an arrow from $Y_i \to Y_j$ unless $f(Y_j|Y_1, \ldots, Y_{j-1})$ does not depend on Y_i, that is unless

$$Y_i \perp\!\!\!\perp Y_j | \{Y_1, \ldots, Y_{j-1}\} \setminus \{Y_i\}$$

When modeling with DAGs, any appropriate univariate response models can be used to specify the conditional densities $f(Y_j|Y_1, \ldots, Y_{j-1})$.

In this part of the thesis we are considering random variables that evolve over time, $Y_{i1}(t), \ldots, Y_{iG}(t)$ and $s_i(t)$. In addition we have the time dependent exogenous covariates $X_1(t), \ldots, X_K(t)$. To incorporate the time factor we give the following definition:

Definition 2.2.2. *A Dynamic directed acyclic graph is defined to be a class of time-indexed DAGs $(\mathcal{G}(t))_{t \geq 0}$ where $\mathcal{G}(t) = (V(t), E(t))$. This is also referred to as a Dynamic path diagram; Fosen et al. (2006).*

In our setting the vertex set can be partitioned into $V(t) = V_c(t) \cup \{s(t)\}$, where $V_c(t)$ is a set of time dependent covariates and $s(t)$ is a binary outcome process of interest. The set of covariates can be further partitioned into $V_c(t) = V_X(t) \cup V_Y(t)$, where $V_X(t) = \{X_1(t), \ldots, X_K(t)\}$ is a set of exogenous covariates, $V_Y(t) = \{Y_1(t), \ldots, Y_G(t)\}$ is a set of endogenous covariates. The directed edge set $E(t)$ may vary with time and we assume

$$E(t) \subset (V_X(t) \times V_Y(t)) \cup (V_Y(t) \times V_Y(t)) \cup (V_c(t) \times \{s(t)\})$$

Edges from an endogenous covariate to an exogenous covariate are not allowed and neither are edges from the binary outcome process to a covariate. We assume that the exogenous covariates are deterministic and implicitly assume that the sub-graph defined by the endogenous covariates, $\{V_Y(t), E_Y(t) = (V_Y(t) \times V_Y(t))\}$, is a DAG at each time point.

The main model under consideration here is the dynamic DAG given in (2.1) or it's latent counterpart (2.2) and the parameters we wish to estimate are the time dependent parameters $\delta_{jk}(t)$, $k = 1, \ldots, d_j$, $j = 1, \ldots, G+1$ and $\gamma_{jk}(t)$, $k = 1, \ldots, g_j$ and $j = 2, \ldots, G+1$. We now introduce the notion of direct, indirect and total effects.

Definition 2.2.3. *If $X_k(t)$ occurs in the jth equation of our dynamic DAG (2.2), or equivalently $Y_j(t)$ is a child of $X_k(t)$, we define the direct effect of $X_k(t)$ on $Y_j(t)$ to be*

$$dir(X_k(t) \to Y_j(t)) = \delta_{jk}(t)$$

Similarly if $Y_i(t)$ occurs in the jth equation, $i < j$, we talk about the direct effect of $Y_i(t)$ on $Y_j(t)$:

$$dir(Y_i(t) \to Y_j(t)) = \gamma_{ji}(t)$$

For the indirect effect of a covariate $X_k(t)$ on $Y_j(t)$ we need to look at all directed paths from $X_k(t)$ to $Y_j(t)$ of length greater than 1 (to distinguish from the direct effect).

A path from $X_k(t)$ will be of the form $\nu_{k,j_1,\ldots,j_m,j} = \{X_k(t), Y_{j_1}, \ldots, Y_{j_m}, Y_j\}$, where $j_1 < \cdots < j_m < j$, and we define the indirect effect of $X_k(t)$ on $Y_j(t)$ mediated through the path $\nu_{k,j_1,\ldots,j_m,j}$ as the product of the direct effects between all adjacent parents and children of the path $\delta_{kj_1}(t)\gamma_{j_1j_2}(t) \cdot \cdots \cdot \gamma_{j_{m-1}j_m}(t)\gamma_{j_mj}(t)$. This leads to the following definition:

Definition 2.2.4. *The indirect effect of $X_k(t)$ on $Y_j(t)$ is*

$$\textit{ind}(X_k(t) \cdots \to Y_j(t)) = \sum_{\nu_{k,j_1,\ldots,j_m,j}} \delta_{kj_1}(t) \left(\prod_{\ell=1}^{m-1} \gamma_{j_\ell j_{\ell+1}}(t) \right) \gamma_{j_m j}(t)$$

where the sum is taken over all directed paths $\nu_{k,j_1,\ldots,j_m,j} = \{X_k(t), Y_{j_1}, \ldots, Y_{j_m}, Y_j\}$ of length greater than 1. Similarly we define the indirect effect of $Y_i(t)$ on $Y_j(t)$, $i < j$

$$\textit{ind}(Y_i(t) \cdots \to Y_j(t)) = \sum_{\nu_{i,j_1,\ldots,j_m,j}} \gamma_{ij_1}(t) \left(\prod_{\ell=1}^{m-1} \gamma_{j_\ell j_{\ell+1}}(t) \right) \gamma_{j_m j}(t)$$

and the sum taken over all directed paths $\nu_{i,j_1,\ldots,j_m,j} = \{Y_i(t), Y_{j_1}, \ldots, Y_{j_m}, Y_j\}$, $i < j_1 < \cdots < j_m < j$.

The total effect is simply the sum of the direct and indirect effect.

Definition 2.2.5. *The total effect of $X_k(t)$ on $Y_j(t)$ is*

$$\textit{tot}(X_k(t) \to Y_j(t)) = \textit{dir}(X_k(t) \to Y_j(t)) + \textit{ind}(X_k(t) \cdots \to Y_j(t))$$

and similarly the total effect of $Y_i(t)$ on $Y_j(t)$ is

$$\textit{tot}(Y_i(t) \to Y_j(t)) = \textit{dir}(Y_i(t) \to Y_j(t)) + \textit{ind}(Y_i(t) \cdots \to Y_j(t))$$

The above effects are of particular interest to the researcher. In particular we are interested in observing how the direct, indirect and total effects, between different variables, change over time. The Bayesian paradigm discussed in the next chapter allows us to estimate all these functional effects and furthermore construct credible bands around them.

2.3 Application in Survival Analysis

In the following section we will show how our model can be used to fit certain types of Survival data. In our construction we will show how under non-informative independent censoring we arrive at model (2.1), where $h_i(t)$ is replaced by the discrete hazard function under interval discretization. The fitting procedure in the Survival Analysis setting is identical to the one described in chapter 3.

Assume we have data arising from a clinical trial where patients, $i = 1, \ldots, n$, enter the study at time 0 and come for follow up visits at pre-specified regular times. At the beginning of the study the experimenter collects some baseline exogenous covariates, such as age, sex or treatment that can be time dependent but not stochastic. This excludes the option of stochastic time-varying treatments, where at time t the treatment depends on the covariate history prior to that time. At each follow up time, including time 0, the experimenter then takes endogenous measurements such as blood pressure, cholesterol level or CD4 cell counts. However, we might have some missing observations and we assume that missingness occurs completely at random. We will address this issue by imputing the missing covariates in our Bayesian fitting scheme of chapter 3. We discretize the time axis with respect to the visiting times into intervals, $[t_0, t_1), [t_1, t_2), \ldots, [t_{m-1}, t_m)$, where $t_0 = 0$ and $t_m = \tau$ denotes the endpoint of the study. We will assume that the failure and censoring times, T_i and C_i are absolutely continuous and conditionally independent given any covariates. However, as with grouped survival data we will not necessarily have information about exact survival times for each individual but rather into which intervals they fall. This serves as an approximation and by letting the number of visits increase and making the interval widths approach 0 we will eventually retrieve the exact failure information. We will also need to make the assumption that the censoring is non-informative in the sense that the

censoring distribution does not depend on the parameters of our model. In our setup the time-points at which measurements are taken for patient i are the same, $t_{ij} = t_j$, for all i. This assumption is required in the application to survival analysis but can be relaxed in the general model setup of section 2.1. We let $\{s_i(t_{ij}), j = 1, \ldots, m_i\}$ denote the binary survival observations of individual i. The value of $s_i(t_{i,j-1})$ at time $t_{i,j-1}$ indicates whether, in the interval $[t_{i,j-1}, t_{i,j}) = [t_{j-1}, t_j)$, individual i survived or was censored , $s_i(t_{i,j-1}) = 0$, or died, $s_i(t_{i,j-1}) = 1$. We will make the assumption that if a patient is at risk at the beginning of interval $[t_{j-1}, t_j)$ the binary variable $s_i(t_{j-1})$ is Bernoulli with death probability

$$h_{i,j} := \mathrm{P}\big[T_i \in [t_{j-1}, t_j) | T_i \geq t_{j-1}, \mathbf{Y}_i(t_{j-1}), \mathbf{X}_i(t_{j-1})\big]$$

where $\mathbf{Y}_i(t_{j-1})$ and $\mathbf{X}_i(t_{j-1})$ denote the endogenous and exogenous vectors of covariates measured at the visiting time t_{j-1}. The above is the discrete hazard of interval $[t_{j-1}, t_j)$ and following the notation of the above section we will model it using the probit model:

$$\Phi^{-1}(h_{i,j}) = \mathbf{Y}_i^{(G+1)}(t_{j-1})\boldsymbol{\gamma}_{G+1}(t_{j-1}) + \mathbf{X}_i^{(G+1)}(t_{j-1})\boldsymbol{\delta}_{G+1}(t_{j-1}) \tag{2.8}$$

This means that the endogenous measurements at the visiting time t_{j-1} along with the exogenous covariates evaluated at t_{j-1} affect whether or not a patient dies in the upcoming time interval preceding his next visit. If we then model the endogenous covariates $Y_{i1}(t), \ldots, Y_{iG}(t)$ according to a dynamic DAG we arrive at the same latent Gaussian model given in (2.1)

$$\begin{aligned}
Y_{i1}(t_{j-1}) &= \mathbf{X}_i^{(1)}(t_{j-1})\boldsymbol{\delta}_1(t_{j-1}) + \varepsilon_1(t_{j-1}) \\
Y_{i2}(t_{j-1}) &= \mathbf{Y}_i^{(2)}(t_{j-1})\boldsymbol{\gamma}_2(t_{j-1}) + \mathbf{X}_i^{(2)}(t_{j-1})\boldsymbol{\delta}_2(t_{j-1}) + \varepsilon_2(t_{j-1}) \\
&\vdots \\
Y_{iG}(t_{j-1}) &= \mathbf{Y}_i^{(G)}(t_{j-1})\boldsymbol{\gamma}_G(t) + \mathbf{X}_i^{(G)}(t_{j-1})\boldsymbol{\delta}_G(t_{j-1}) + \varepsilon_G(t_{j-1}) \\
\Phi^{-1}(h_{i,j}) &= \mathbf{Y}_i^{(G+1)}(t_{j-1})\boldsymbol{\gamma}_{G+1}(t_{j-1}) + \mathbf{X}_i^{(G+1)}(t_{j-1})\boldsymbol{\delta}_{G+1}(t_{j-1})
\end{aligned} \tag{2.9}$$

Sometimes we might want to replace the left endpoint, t_{j-1}, of our interval in the above equations with another point, $t^*_{j-1} \in (t_{j-1}, t_j)$, say the midpoint of the interval. This is a correction often made in lifetime analysis and could for example be applied to the last equation. The parameters of this model can now be fitted using the Bayesian estimation scheme of chapter 3. A thorough justification of this model, especially in terms of censoring, can be found in the appendix. We will end this section with a couple of remarks about the model given in (2.9).

Remark 2.3.1. *An important thing to note is that in the $G + 1$ equation of the system (2.9) we are assuming that the covariates affect the discrete hazard in the intervals $[t_{j-1}, t_j)$ in a smooth way. This is unrealistic if the interval lengths $t_j - t_{j-1}$ are not all equal. Also note that we might be more interested in estimating the continuous hazard function, $h_i(t_{j-1})$ but not the discrete probabilities $h_{i,j}$. It turns out that we can resolve this issue and get an approximate model for the continuous hazard by including an offset on the right hand side of (2.8). The idea of using an offset in the above context is motivated by Efron (1988) who used an offset in a similar setting when analyzing survival data using logistic regression. In order to estimate the continuous hazard we use the approximation $h_{i,j} = h_i(t_{j-1})\Delta_j$, where Δ_j is the length of the jth interval, $[t_{j-1}, t_j)$. The idea is then to approximate the probit link with a stretched logit link and include an offset term of the form $K \log \Delta_j$ for some constant K. This way, as we will see below, the offset will cancel with a corresponding term on the left side of the probit model giving rise to an approximate estimate of $\Phi^{-1}(h_i(t_{j-1})$. To be more specific we approximate the cumulative normal density in the following way*

$$\Phi(s) \approx \frac{e^{cs}}{1 + e^{cs}} \tag{2.10}$$

where c is chosen so as to make the approximation as accurate as possible. Demidenko (2004) explains how using numerical computation, one can show that with a value of

$c \approx 1.7$ *this approximation has the following minimax accuracy*

$$\max_{-\infty<s<\infty} \left| \Phi(s) - \frac{e^{cs}}{1+e^{cs}} \right| = 0.00946$$

By inverting (2.10) *we reach the following approximation*

$$\Phi^{-1}(s) \approx \frac{1}{c} \log\left(\frac{s}{1-s}\right) = 0.6 \log\left(\frac{s}{1-s}\right)$$

evaluated at $c = 1.7$. *This suggest using the offset term* $0.6 \log \Delta_j$ *under the approximation* $\log \frac{p}{1-p} = \log p$ *for small* p. *Note that we are using two approximations here, one involves approximating the probit with a logit and the second involves approximating the logit with a* $\log$. *The model that should be fitted is the following*

$$\Phi^{-1}\left(h_{i,j}\right) \;=\; 0.6 \log \Delta_j + \mathbf{Y}_i^{(G+1)}(t_{j-1})\boldsymbol{\gamma}_{G+1}(t_{j-1}) + \mathbf{X}_i^{(G+1)}(t_{j-1})\boldsymbol{\delta}_{G+1}(t_{j-1})$$

We see that the above approximations lead to

$$\begin{aligned}
\Phi^{-1}\left(h_i(t_{j-1})\right) &\approx 0.6 \log\left(\frac{h_i(t_{j-1})}{1-h_i(t_{j-1})}\right) \\
&\approx 0.6 \log\left(h_i(t_{j-1})\right) \\
&\approx 0.6 \log\left(h_{i,j}\right) - 0.6 \log \Delta_j \\
&\approx \Phi^{-1}\left(h_{i,j}\right) - 0.6 \log \Delta_j
\end{aligned}$$

but this shows that by fitting the above probit model, with the offset, we end up with an approximate estimate of the continuous hazard

$$\hat{h}(t_{j-1}) = \Phi\left(\mathbf{Y}_i^{(G+1)}(t_{j-1})\hat{\boldsymbol{\gamma}}_{G+1}(t_{j-1}) + \mathbf{X}_i^{(G+1)}(t_{j-1})\hat{\boldsymbol{\delta}}_{G+1}(t_{j-1})\right)$$

Remark 2.3.2. *Another important point is that the logistic regression model for the discrete hazard given in Efron (1988) is directly related to a special case of the model* (2.9). *The model he considered involved only the last equation, had no subject specific covariates and fixed covariate effects only*

$$\log\left(\frac{h_j}{1-h_j}\right) = \mathbf{W}(t_{j-1})\boldsymbol{\beta}$$

where the $\boldsymbol{\beta}$ is assumed fixed and $h_{i,j} = h_j$, for all $i = 1, \ldots, n$. By replacing the logit link with the probit approximation this becomes a special case of our proposed model. The likelihood of our data, which we derive more formally in the appendix, takes on the form

$$L = \prod_{j=1}^{m} \prod_{i=1}^{n} h_j^{d_{ij}} (1 - h_j)^{r_{ij} - d_{ij}}$$

where r_{ij} is the at risk indicator denoting whether patient i is at risk at the beginning of interval j and d_{ij} records whether patient i dies during the jth interval. But the above translates, up to a scaling factor, directly into the binomial likelihood used in Efron (1988)

$$L = \prod_{j=1}^{m} \binom{n_j}{s_j} h_j^{s_j} (1 - h_j)^{n_j - s_j}$$

where $n_j = \sum_i r_{ij}$ is the number of patients at risk at the beginning of interval j and $s_j = \sum_i d_{ij}$ is the number of deaths in interval j.

CHAPTER 3

ESTIMATION

In this chapter we will construct an MCMC Gibbs sampler for the estimation of the parameters in model (2.7) both under independent errors and AR(1) errors. The unobserved latent variables $Y_{i,G+1}(t_{ij})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, will be imputed at each iteration step of the Gibbs sampler just like is done in problems of missing data. Before continuing we will introduce the following notation. Let $\mathbf{s}_i = (s_i(t_{i1}), \ldots, s(t_{im_i}))'$ be the binary outcome vector for individual i, and let $\mathbf{s} = (\mathbf{s}_1', \ldots, \mathbf{s}_n')'$. Define $\mathbf{y}_{i,obs}(t_{ij}) := (y_{i1}(t_{ij}), \ldots, y_{iG}(t_{ij}))'$. Then the complete data vector from (2.5) is $\mathbf{y}_i(t_{ij}) = (\mathbf{y}_{i,obs}(t_{ij})', y_{i,G+1}(t_{ij}))'$. Let $\mathbf{y}_{i,obs}$ and $\mathbf{y}_{i,G+1}$ denote the vectors consisting of $\mathbf{y}_{i,obs}(t_{ij})$ and $y_{i,G+1}(t_{ij})$ respectively for all t_{ij}, $j = 1, \ldots, m_i$. Similarly we define $\mathbf{y}_{obs}$ and $\mathbf{y}_{G+1}$ as the complete observed response vector and the complete latent response vector. Finally we denote by $\mathbf{y}_i$ and $\mathbf{y}$ the complete response vectors from models (2.6) and (2.7) respectively.

3.1 Independence across time

Let us consider the fitting procedures under independent covariance structure. Since we both have independence across individuals and across time it will prove most convenient to work with the model specification in (2.5)

$$\mathbf{Y}_i(t_{ij}) = \mathbf{W}_i(t_{ij})\boldsymbol{\beta} + \mathbf{Z}_i(t_{ij})\mathbf{u} + \boldsymbol{\varepsilon}_i(t_{ij})$$

with $\boldsymbol{\varepsilon}_i(t_{ij}) \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, all $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. We recall that $\boldsymbol{\Sigma} = \mathrm{diag}_{1 \le k \le G+1}(\sigma_k^2)$ and $\sigma_{G+1}^2 = 1$ because of issues with identifiability. We place the

following priors on the parameters

$$\begin{aligned}
[\boldsymbol{\beta}] &\equiv 1 \\
\mathbf{u} &\sim \mathrm{N}(\mathbf{0}, \mathbf{D}_u \otimes \mathbf{I}_\kappa) \\
\sigma^2_{u\ell} &\overset{iid}{\sim} \mathrm{IG}(A_u, B_u), \quad \ell = 1, \ldots, g+d \\
\sigma^2_k &\overset{iid}{\sim} \mathrm{IG}(A_k, B_k), \quad k = 1, \ldots, G \\
\sigma^2_{G+1} &= 1
\end{aligned} \tag{3.1}$$

where as we recall $\mathbf{D}_u = \mathrm{diag}_{1 \le \ell \le g+d}(\sigma^2_{u\ell})$ and IG denotes the inverse gamma distribution. To make the priors on $\sigma^2_{u\ell}$ and σ^2_k as non-informative as possible we choose A_u, B_u, A_k and B_k to be close to zero.

Let $\boldsymbol{\Theta}$ denote the set of parameters $\{\boldsymbol{\beta}, \mathbf{u}, (\sigma^2_{u\ell})_{\ell=1,\ldots,g+d}, \boldsymbol{\Sigma}\}$. Then except for a constant of proportionality, the posterior density is equal to

$$\big[\boldsymbol{\Theta}, \mathbf{y}_{G+1} | \mathbf{y}_{obs}, \mathbf{s}\big] \;\propto\; \big[\mathbf{s} | \mathbf{y}_{G+1}, \mathbf{y}_{obs}, \boldsymbol{\Theta}\big] \big[\mathbf{y} | \boldsymbol{\Theta}\big] \big[\boldsymbol{\Theta}\big] \tag{3.2}$$

with

$$\big[\mathbf{y} | \boldsymbol{\Theta}\big] = \prod_{i=1}^{n} \prod_{j=1}^{m_i} \big[\mathbf{y}_i(t_{ij}) | \boldsymbol{\Theta}\big], \tag{3.3}$$

$$\big[\mathbf{s} | \mathbf{y}_{G+1}, \mathbf{y}_{obs}, \boldsymbol{\Theta}\big] = \prod_{i=1}^{n} \prod_{j=1}^{m_i} \mathrm{I}(y_{i,G+1}(t_{ij}) \in A_{ij}(s_i(t_{ij})) \tag{3.4}$$

$$\big[\boldsymbol{\Theta}\big] = \big[\boldsymbol{\beta}\big] \big[\mathbf{u} | \sigma^2_{u1}, \ldots, \sigma^2_{u,g+d}\big] \prod_{\ell=1}^{g+d} \big[\sigma^2_{u\ell}\big] \prod_{k=1}^{G+1} \big[\sigma^2_k\big]$$

where we define $A_{ij}(s_i(t_{ij})) := \{y_{i,G+1}(t_{ij}) | y_{i,G+1}(t_{ij}) > 0 \text{ if } s_i(t_{ij}) = 1$ $\& \; y_{i,G+1}(t_{ij}) \le 0 \text{ if } s_i(t_{ij}) = 0\}$. The equality in (3.4) will be explained below but let us now explain why the vectors $\mathbf{y}_i(t_{ij})$ are indeed independent leading us to the product formula in (3.3). In the process we will derive a formula for the density that will prove useful to us when constructing the Gibbs sampler for the parameters. If we look back

at the system of equations in (2.2) we note that this is a typical simultaneous equation system that can be written in the following way

$$\mathbf{Y}_i(t) = \mathbf{\Gamma}(t)\mathbf{Y}_i(t) + \mathbf{\Delta}(t)\mathbf{X}_i(t) + \boldsymbol{\varepsilon}_i(t)$$

where $\mathbf{\Gamma}(t)$ is a $(G+1)\times(G+1)$ matrix whose elements are the components of $\boldsymbol{\gamma}_j(t) = (\gamma_{j1}(t), \ldots, \gamma_{jg_j}(t))'$, $2 \leq j \leq G+1$, and $\mathbf{\Delta}(t)$ is a $(G+1)\times K$ matrix whose elements are the components of $\boldsymbol{\delta}_j(t) = (\delta_{j1}(t), \ldots, \delta_{jd_j}(t))'$, $1 \leq j \leq G+1$. Moving the term involving $\mathbf{Y}_i(t)$ to the left side of the equation leads to

$$(\mathbf{I} - \mathbf{\Gamma}(t))\mathbf{Y}_i(t) = \mathbf{\Delta}(t)\mathbf{X}_i(t) + \boldsymbol{\varepsilon}_i(t) \tag{3.5}$$

and note that since we are dealing with a directed acyclic graph the matrix $\mathbf{\Gamma}(t)$ is lower triangular with zeros on the diagonal and hence $\mathbf{I} - \mathbf{\Gamma}(t)$ is invertible. Multiplying through the above equation by $(\mathbf{I} - \mathbf{\Gamma}(t))^{-1}$ we establish that

$$\mathbf{Y}_i(t) \sim \mathrm{N}((\mathbf{I} - \mathbf{\Gamma}(t))^{-1}\mathbf{\Delta}(t)\mathbf{X}_i(t), (\mathbf{I} - \mathbf{\Gamma}(t))^{-1}\mathbf{\Sigma}(\mathbf{I} - \mathbf{\Gamma}(t))^{-T}) \tag{3.6}$$

Since $\boldsymbol{\varepsilon}_i(t_{ij})$ are independent for all i, j, it follows that $(\mathbf{I} - \mathbf{\Gamma}(t_{ij}))^{-1}\boldsymbol{\varepsilon}_i(t_{ij})$ and hence $\mathbf{Y}_i(t_{ij})$ are independent for all i, j. From (3.6) we arrive at the density of $\mathbf{Y}_i(t)$:

$$\begin{aligned} & (2\pi)^{-(G+1)/2}|\mathbf{\Sigma}|^{-1/2}\exp\Big(-\frac{1}{2}\left(\mathbf{Y}_i(t) - (\mathbf{I} - \mathbf{\Gamma}(t))^{-1}\mathbf{\Delta}(t)\mathbf{X}_i(t)\right)' \\ & \times\left[(\mathbf{I} - \mathbf{\Gamma}(t))'\mathbf{\Sigma}^{-1}(\mathbf{I} - \mathbf{\Gamma}(t))\right]\left(\mathbf{Y}_i(t) - (\mathbf{I} - \mathbf{\Gamma}(t))^{-1}\mathbf{\Delta}(t)\mathbf{X}_i(t)\right)\Big) \\ = \quad & (2\pi)^{-(G+1)/2}|\mathbf{\Sigma}|^{-1/2}e^{-\frac{1}{2}(\mathbf{Y}_i(t) - \mathbf{\Gamma}(t)\mathbf{Y}_i(t) - \mathbf{\Delta}(t)\mathbf{X}_i(t))'\mathbf{\Sigma}^{-1}(\mathbf{Y}_i(t) - \mathbf{\Gamma}(t)\mathbf{Y}_i(t) - \mathbf{\Delta}(t)\mathbf{X}_i(t))} \end{aligned}$$

Define $\boldsymbol{\mu}_i(t_{ij}) := \mathbf{W}_i(t_{ij})\boldsymbol{\beta} + \mathbf{Z}_i(t_{ij})\mathbf{u}$ which by the construction in the previous section equals $\mathbf{\Gamma}(t_{ij})\mathbf{Y}_i(t_{ij}) + \mathbf{\Delta}(t_{ij})\mathbf{X}_i(t_{ij})$. Then by the above derivation we've established the following working formula for the density of $\mathbf{Y}_i(t_{ij})$

$$\left[\mathbf{y}_i(t_{ij})|\mathbf{\Theta}\right] = (2\pi)^{-(G+1)/2}|\mathbf{\Sigma}|^{-1/2}e^{-\frac{1}{2}(\mathbf{y}_i(t_{ij}) - \boldsymbol{\mu}_i(t_{ij}))'\mathbf{\Sigma}^{-1}(\mathbf{y}_i(t_{ij}) - \boldsymbol{\mu}_i(t_{ij}))} \tag{3.7}$$

Let us now derive the equality in (3.4). Since $\mathbf{y}_i(t_{ij})$ are independent for all i, j we have that $s_i(t_{ij}) = \mathrm{I}(Y_{i,G+1}(t_{ij}) > 0)$ are independent for all i, j which leads to the following

$$\begin{aligned}
\big[\mathbf{s}|\mathbf{y}_{G+1}, \mathbf{y}_{obs}, \boldsymbol{\Theta}\big] &= \big[\big(s_i(t_{ij})\big)_{i,j}|\mathbf{y}_{G+1}, \mathbf{y}_{obs}, \boldsymbol{\Theta}\big] \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m_i}\big[s_i(t_{ij})|\mathbf{y}_{G+1}, \mathbf{y}_{obs}, \boldsymbol{\Theta}\big] \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m_i}\big[s_i(t_{ij})|y_{i,G+1}(t_{ij})\big] \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m_i}\mathrm{I}(y_{i,G+1}(t_{ij}) > 0)^{s_i(t_{ij})}\mathrm{I}(y_{i,G+1}(t_{ij}) \leq 0)^{1-s_i(t_{ij})} \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m_i}\mathrm{I}(y_{i,G+1}(t_{ij}) \in A_{ij}(s_i(t_{ij}))
\end{aligned}$$

3.1.1 Gibbs sampler for the parameters

The conditional posterior of $(\boldsymbol{\beta}, \mathbf{u})$ given $(\sigma^2_{u1}, \ldots, \sigma^2_{u,g+d}, \boldsymbol{\Sigma}, \mathbf{y}_{G+1})$ can be derived by looking at the terms on the right hand side of (3.2) that involve $\boldsymbol{\beta}$ and $\mathbf{u}$. It follows from (3.3) and (3.7) that

$$\big[\mathbf{y}|\boldsymbol{\Theta}\big] = (2\pi)^{-(G+1)(\sum_{i=1}^{n} m_i)/2}|\boldsymbol{\Omega}|^{-1/2}e^{-\frac{1}{2}(\mathbf{y}-\mathbf{W}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{W}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})}$$

where we recall $\boldsymbol{\Omega} = \mathbf{I} \otimes \boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{\varepsilon}$ in (2.7). From this it is easy to see that the complete conditional of $(\boldsymbol{\beta}, \mathbf{u})$ is proportional to

$$\exp\left\{-\frac{1}{2}\left[(\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}'(\mathbf{D}_u^{-1} \otimes \mathbf{I}_\kappa)\mathbf{u}\right]\right\}$$

By the usual technique of completing the square it can be shown that

$$(\boldsymbol{\beta}, \mathbf{u})'|\mathbf{y}, \mathbf{s}, \boldsymbol{\Theta}\backslash\{\boldsymbol{\beta}, \mathbf{u}\} \;\sim\; \mathbf{N}(\boldsymbol{\mu}_{\beta,u}, \boldsymbol{\Sigma}_{\beta,u}) \tag{3.8}$$

where we define

$$\begin{aligned}
\boldsymbol{\mu}_{\beta,u} &= (\mathbf{M}'\boldsymbol{\Omega}^{-1}\mathbf{M} + \mathbf{D})^{-1}\mathbf{M}'\boldsymbol{\Omega}^{-1}\mathbf{y} \\
\boldsymbol{\Sigma}_{\beta,u} &= (\mathbf{M}'\boldsymbol{\Omega}^{-1}\mathbf{M} + \mathbf{D})^{-1}
\end{aligned}$$

with $\mathbf{M} = [\mathbf{W} \quad \mathbf{Z}]$ and $\mathbf{D} = \text{blockdiag}(\mathbf{0}, \mathbf{D}_u^{-1} \otimes \mathbf{I}_\kappa)$, see Ruppert et al. (2003), chapter 16 for details. Using the blockdiagonal structure of $\mathbf{\Omega}$ and $\mathbf{D}$ the mean and covariance above are easily computed.

To derive the complete conditional distribution for $\sigma^2_{u\ell}$ we note that the prior distribution is inverse gamma

$$\left[\sigma^2_{u\ell}\right] = \frac{B_u^{A_u}}{\Gamma(A_u)}(\sigma^2_{u\ell})^{-(A_u+1)} \exp\left(-\frac{B_u}{\sigma^2_{u\ell}}\right)$$

and the density of $\mathbf{u}|\sigma^2_{u1}, \ldots, \sigma^2_{u,g+d}$ is

$$(2\pi)^{-\kappa(g+d)/2}\left(\prod_{\ell=1}^{g+d}(\sigma^2_{u\ell})^{\kappa}\right)^{-1/2} \exp\left(-\frac{1}{2}\sum_{\ell=1}^{g+d}\frac{\|\mathbf{u}_\ell\|^2}{\sigma^2_{u\ell}}\right)$$

so the complete conditional for $\sigma^2_{u\ell}$, for all $\ell = 1, \ldots, g+d$, is proportional to

$$(\sigma^2_{u\ell})^{-(A_u+\frac{\kappa}{2}+1)} \exp\left\{-\frac{1}{\sigma^2_{u\ell}}\Big(B_u + \frac{1}{2}\|\mathbf{u}_\ell\|^2\Big)\right\}$$

We thus need to sample independently from

$$\sigma^2_{u\ell}|\mathbf{y}, \mathbf{s}, \mathbf{\Theta}\backslash\{\sigma^2_{u\ell}\} \quad \sim \quad \text{IG}\left(A_u + \frac{\kappa}{2}, B_u + \frac{1}{2}\|\mathbf{u}_\ell\|^2\right) \tag{3.9}$$

By exploring (3.2) we note that the complete conditional of σ^2_k, $k = 1, \ldots, G$ is proportional to

$$\left[\sigma^2_k\right]\left[\mathbf{y}|\mathbf{\Theta}\right]$$

where

$$\left[\sigma^2_k\right] = \frac{B_k^{A_k}}{\Gamma(A_k)}(\sigma^2_k)^{-(A_k+1)} \exp\left(-\frac{B_k}{\sigma^2_k}\right) \tag{3.10}$$

and the complete data density is proportional to

$$|\mathbf{\Sigma}|^{-\sum_{i=1}^n m_i/2} \exp\left(-\frac{1}{2}\sum_{i,j}(\mathbf{y}_i(t_{ij}) - \boldsymbol{\mu}_i(t_{ij}))'\mathbf{\Sigma}^{-1}(\mathbf{y}_i(t_{ij}) - \boldsymbol{\mu}_i(t_{ij}))\right)$$

Focusing only on terms that involve σ_k^2 and recalling that $m := \sum_{i=1}^n m_i$ the complete data density can be shown to be proportional to

$$(\sigma_k^2)^{-m/2} \exp\left(-\frac{1}{2} \sum_{i,j} \frac{\big(y_{ik}(t_{ij}) - \mu_{ik}(t_{ij})\big)^2}{\sigma_k^2} \right) \tag{3.11}$$

where we recall $\boldsymbol{\mu}_i(t_{ij}) := \mathbf{W}_i(t_{ij})\boldsymbol{\beta} + \mathbf{Z}_i(t_{ij})\mathbf{u}$ and we let $\mu_{ik}(t_{ij})$ denote the kth component of the vector. Combining (3.10) and (3.11) we arrive at the complete conditional of σ_k^2, $k = 1, \ldots, G$:

$$\sigma_k^2 | \mathbf{y}, \mathbf{s}, \boldsymbol{\Theta} \backslash \{\sigma_k^2\} \sim \text{IG}\Big(A_k + \frac{m}{2}, B_k + \frac{1}{2} \sum_{i,j} \big(y_{ik}(t_{ij}) - \mu_{ik}(t_{ij})\big)^2 \Big)$$

3.1.2 Imputing the data

Imputing the latent variable

When imputing the data $\mathbf{y}_{G+1}$ we need to find the complete conditional distribution of $\mathbf{y}_{G+1}$ given the observed data and the parameters of the model. We isolate the parts on the right hand side of (3.2) that depend on $\mathbf{y}_{G+1}$ and find that

$$\begin{aligned} \big[\mathbf{y}_{G+1} | \mathbf{y}_{obs}, \mathbf{s}, \boldsymbol{\Theta}\big] &\propto \prod_{i=1}^n \prod_{j=1}^{m_i} \big[\mathbf{y}_i(t_{ij}) | \boldsymbol{\Theta}\big] \prod_{i=1}^n \prod_{j=1}^{m_i} \mathrm{I}(y_{i,G+1}(t_{ij}) \in A_{ij}(s_i(t_{ij})) \\ &\propto \prod_{i=1}^n \prod_{j=1}^{m_i} \big[y_{i,G+1}(t_{ij}) | \mathbf{y}_{i,obs}(t_{ij}), \boldsymbol{\Theta}\big] \mathrm{I}(y_{i,G+1}(t_{ij}) \in A_{ij}(s_i(t_{ij})) \end{aligned}$$

This shows that in order to sample from the posterior distribution of $\mathbf{y}_{G+1}$, given the parameters, we need to sample independently from truncated normal distributions. Given the parameters $(\boldsymbol{\beta}, \mathbf{u})'$ we can directly calculate $\boldsymbol{\gamma}_{G+1}(t_{ij})$ and $\boldsymbol{\delta}_{G+1}(t_{ij})$ in (2.2) and hence we simulate the latent variable based on a truncated normal with mean $\mu_{i,G+1}(t_{ij}) = \mathbf{Y}_i^{(G+1)}(t_{ij})\boldsymbol{\gamma}_{G+1}(t_{ij}) + \mathbf{X}_i^{(G+1)}(t_{ij})\boldsymbol{\delta}_{G+1}(t_{ij})$ and variance 1.

Imputing missing covariates

If a patient i does not show up for a follow up visit at time t_{ij}, but is observed at a later time, or exact time of death is known to be in a later interval, we know the patient is alive at time t_{ij} but won't have any information about the first G endogenous covariates. Hence at that time-point the vector $\mathbf{y}_i(t_{ij}) = (y_{i1}(t_{ij}), \ldots, y_{iG}(t_{ij}), y_{i,G+1}(t_{ij}))$ is completely unobserved. These missing data, which are assumed to be missing completely at random, are easily handled in our Bayesian framework by imputing the unobserved data vector as a block at each iteration. Since the errors for different individuals and times are independent, it all boils down to simulating independently $\mathbf{y}_i(t_{ij})|\boldsymbol{\Theta}$ from the normal distribution in (3.6) truncated with $\mathrm{I}(y_{i,G+1}(t_{ij}) \in A_{ij}(s_i(t_{ij}))$. We emphasize again that the matrices $\boldsymbol{\Gamma}(t_{ij})$ and $\boldsymbol{\Delta}(t_{ij})$ are easily computed with knowledge of the parameter values $(\boldsymbol{\beta}, \mathbf{u})'$.

3.2 Time dependence

In this section we will consider the exact same model as in the previous section except that we will impose a different covariance structure on the errors. Let us consider the k-th equation of $(5!)$,

$$Y_{ik}(t) = \mathbf{W}_{ik}(t)\boldsymbol{\beta} + \mathbf{Z}_{ik}(t)\mathbf{u} + \varepsilon_{ik}(t), \tag{3.12}$$

where $\mathbf{W}_{ik}(t)$ and $\mathbf{Z}_{ik}(t)$ are defined to be the the k-th rows of the matrices $\mathbf{W}_i(t)$ and $\mathbf{Z}_i(t)$, $Y_{ik}(t)$ represents the k-th component of $\mathbf{Y}_i(t)$ and $\varepsilon_{ik}(t)$ the corresponding error term. Individual i is observed at time points $t_{i1}, \ldots, t_{im_i}$ and we wish to model the correlation structure of $(\varepsilon_{ik}(t_{i1}), \ldots, \varepsilon_{ik}(t_{im_i}))'$, for $k = 1, \ldots, G+1$. We assume that the time dependent error follows a stationary $\mathrm{AR}(1)$ process and the error at time t_{i1}

arises from the stationary distribution:

$$\begin{aligned}\varepsilon_{ik}(t_{i1}) &= e_{ik}(t_{i1}) \sim \mathbf{N}(0, \sigma_k^2/(1-\phi_k^2)) \\ \varepsilon_{ik}(t_{ij}) &= \phi_k \varepsilon_{ik}(t_{i,j-1}) + e_{ik}(t_{ij}), \quad j \geq 2 \end{aligned} \tag{3.13}$$

and $e_{ik}(t_{ij}) \overset{iid}{\sim} \mathbf{N}(0, \sigma_k^2)$ for all $j \geq 2$. Furthermore $e_{ik}(t_{ij})$ are independent for all $i = 1, \ldots, n$, $j = 1, \ldots, m_i$ and all $k = 1, \ldots, G+1$. This introduces $G+1$ new parameters, $(\phi_1, \ldots, \phi_{G+1})$, into the model. It is easy to see that by using the above error structure and by plugging into (3.12) we get

$$Y_{ik}^*(t_{ij}) = \mathbf{W}_{ik}^*(t_{ij})\boldsymbol{\beta} + \mathbf{Z}_{ik}^*(t_{ij})\mathbf{u} + e_{ik}(t_{ij}), \tag{3.14}$$

for all $j = 1, \ldots, m_i$, where we have defined

$$\begin{aligned} Y_{ik}^*(t_{i1}) &:= Y_{ik}(t_{i1}) \\ Y_{ik}^*(t_{ij}) &:= Y_{ik}(t_{ij}) - \phi_k Y_{ik}(t_{i,j-1}), \quad j \geq 2 \\ \mathbf{W}_{ik}^*(t_{i1}) &:= \mathbf{W}_{ik}(t_{i1}) \\ \mathbf{W}_{ik}^*(t_{ij}) &:= \mathbf{W}_{ik}(t_{ij}) - \phi_k \mathbf{W}_{ik}(t_{i,j-1}), \quad j \geq 2 \\ \mathbf{Z}_{ik}^*(t_{i1}) &:= \mathbf{Z}_{ik}(t_{i1}) \\ \mathbf{Z}_{ik}^*(t_{ij}) &:= \mathbf{Z}_{ik}(t_{ij}) - \phi_k \mathbf{Z}_{ik}(t_{i,j-1}), \quad j \geq 2 \end{aligned}$$

Define $\mathbf{Y}_i^*(t_{ij}) := (Y_{i1}^*(t_{ij}), \ldots, Y_{i,G+1}^*(t_{ij}))'$, $\mathbf{e}_i(t_{ij}) := (e_{i1}(t_{ij}), \ldots, e_{i,G+1}(t_{ij}))'$ and let $\mathbf{W}_i^*(t_{ij})$ and $\mathbf{Z}_i^*(t_{ij})$ be the matrices whose k-th rows are $\mathbf{W}_{ik}^*(t_{ij})$ and $\mathbf{Z}_{ik}^*(t_{ij})$ respectively. Then (3.14) can be expressed in vector form as

$$\mathbf{Y}_i^*(t_{ij}) = \mathbf{W}_i^*(t_{ij})\boldsymbol{\beta} + \mathbf{Z}_i^*(t_{ij})\mathbf{u} + \mathbf{e}_i(t_{ij}), \tag{3.15}$$

for all $j = 1, \ldots, m_i$. Define $\boldsymbol{\Sigma}_1 := \text{diag}_{1 \leq k \leq G+1}\left(\sigma_k^2/(1-\phi_k^2)\right)$ and $\boldsymbol{\Sigma}_j := \text{diag}_{1 \leq k \leq G+1}\left(\sigma_k^2\right)$, for $j = 2, \ldots, m_i$. These matrices correspond to the covariance matrices of the error vectors, $\mathbf{e}_i(t_{ij})$, for all j. In the same way as in the previous chapter

we stack the vectors $\mathbf{Y}_i^*(t_{ij})$ and $\mathbf{e}_i(t_{ij})$ and the matrices $\mathbf{W}_i^*(t_{ij})$ and $\mathbf{Z}_i^*(t_{ij})$ on top of one another, first by the observed times, $t_{i1}, \ldots, t_{im_i}$, and then by the individuals $i = 1, \ldots, n$. This will lead to the following two representations

$$\mathbf{Y}_i^* = \mathbf{W}_i^* \boldsymbol{\beta} + \mathbf{Z}_i^* \mathbf{u} + \mathbf{e}_i, \tag{3.16}$$

and

$$\mathbf{Y}^* = \mathbf{W}^* \boldsymbol{\beta} + \mathbf{Z}^* \mathbf{u} + \mathbf{e}, \tag{3.17}$$

with $\boldsymbol{\Omega}_i := \text{Cov}(\mathbf{e}_i) = \text{blockdiag}_{1 \le j \le m_i}(\boldsymbol{\Sigma}_j)$ and $\boldsymbol{\Omega} := \text{Cov}(\mathbf{e}) = \text{blockdiag}(\boldsymbol{\Omega}_i)$. The vectors and matrices in the above two models are defined in the natural way and analogously to what was done in the previous chapter.

3.2.1 Posterior calculations for the autocorrelated case

We place the same priors on the parameters as before

$$\begin{aligned}
[\boldsymbol{\beta}] &\equiv 1 \\
\mathbf{u} &\sim \text{N}(\mathbf{0}, \mathbf{D}_u \otimes \mathbf{I}_\kappa) \\
\sigma_{u\ell}^2 &\overset{iid}{\sim} \text{IG}(A_u, B_u), \quad \ell = 1, \ldots, g+d \\
\sigma_k^2 &\overset{iid}{\sim} \text{IG}(A_k, B_k), \quad k = 1, \ldots, G \\
\sigma_{G+1}^2 &= 1
\end{aligned} \tag{3.18}$$

and in addition assume a uniform prior on the new parameters concentrated on the interval that guarantees stationarity of the AR(1) process

$$[\phi_k] = \mathbf{I}_{(-1,1)}(\phi_k), \quad k = 1, \ldots, G+1 \tag{3.19}$$

where I denotes the indicator function. As before we let $\boldsymbol{\Theta} := \{\boldsymbol{\beta}, \mathbf{u},$ $(\sigma_{u\ell}^2)_{\ell=1,\ldots,g+d}, (\sigma_k^2)_{k=1,\ldots,G+1}, (\phi_k)_{k=1,\ldots,G+1}\}$ denote the set of parameters. Then as in

the independent case except for a constant of proportionality, the posterior density is equal to

$$\left[\boldsymbol{\Theta}, \mathbf{y}_{G+1} | \mathbf{y}_{obs}, \mathbf{s}\right] \quad \propto \quad \left[\mathbf{s} | \mathbf{y}_{G+1}, \mathbf{y}_{obs}, \boldsymbol{\Theta}\right] \left[\mathbf{y} | \boldsymbol{\Theta}\right] \left[\boldsymbol{\Theta}\right] \tag{3.20}$$

where

$$\left[\boldsymbol{\Theta}\right] \quad = \quad \left[\boldsymbol{\beta}\right] \left[\mathbf{u} | \sigma^2_{u1}, \ldots, \sigma^2_{u,g+d}\right] \prod_{\ell=1}^{g+d} \left[\sigma^2_{u\ell}\right] \prod_{k=1}^{G+1} \left[\sigma^2_k\right] \prod_{k=1}^{G+1} \left[\phi_k\right]$$

and we have the two following lemmas

Lemma 3.2.1. *Let* $\mathbf{y}^*, \mathbf{W}^*, \mathbf{Z}^*$ *be defined as in* (3.17) *and* $\boldsymbol{\Omega}$ *the covariance matrix of the stacked errors,* $\mathbf{e}$*, as defined above. Then the likelihood of* $\mathbf{y}$ *is*

$$\begin{aligned} \left[\mathbf{y} | \boldsymbol{\Theta}\right] \quad &= \quad \prod_{i=1}^{n} \left(\left[\mathbf{y}_i(t_{i1}) | \boldsymbol{\Theta}\right] \prod_{j=2}^{m_i} \left[\mathbf{y}_i(t_{ij}) | \mathbf{y}_i(t_{i,j-1}), \boldsymbol{\Theta}\right] \right) \\ &= \quad (2\pi)^{-(G+1)m/2} |\boldsymbol{\Omega}|^{-1/2} e^{-\frac{1}{2}(\mathbf{y}^* - \mathbf{W}^*\boldsymbol{\beta} - \mathbf{Z}^*\mathbf{u})' \boldsymbol{\Omega}^{-1} (\mathbf{y}^* - \mathbf{W}^*\boldsymbol{\beta} - \mathbf{Z}^*\mathbf{u})} \end{aligned} \tag{3.21}$$

Proof: See appendix.

Lemma 3.2.2. *As in the independent case we have the following result*

$$\left[\mathbf{s} | \mathbf{y}_{G+1}, \mathbf{y}_{obs}, \boldsymbol{\Theta}\right] \quad = \quad \prod_{i=1}^{n} \prod_{j=1}^{m_i} I(y_{i,G+1}(t_{ij}) \in A_{ij}(s_i(t_{ij})) \tag{3.22}$$

where $A_{ij}(s_i(t_{ij})) := \{y_{i,G+1}(t_{ij}) | y_{i,G+1}(t_{ij}) > 0$ *if* $s_i(t_{ij}) = 1$

$\&\ y_{i,G+1}(t_{ij}) \leq 0$ *if* $s_i(t_{ij}) = 0\}$.

Proof: See appendix.

3.2.2 Gibbs sampler for the parameters

We yield the complete conditional of $(\boldsymbol{\beta}, \mathbf{u})$ immediately by using the likelihood in (3.21) and combining it with the prior on $(\boldsymbol{\beta}, \mathbf{u})'$:

$$(\boldsymbol{\beta}, \mathbf{u})' | \mathbf{y}, \mathbf{s}, \boldsymbol{\Theta} \backslash \{\boldsymbol{\beta}, \mathbf{u}\} \sim \mathbf{N}(\boldsymbol{\mu}_{\beta,u}, \boldsymbol{\Sigma}_{\beta,u})$$

where

$$\begin{aligned}\boldsymbol{\mu}_{\beta,u} &= (\mathbf{M}^{*'}\boldsymbol{\Omega}^{-1}\mathbf{M}^{*}+\mathbf{D})^{-1}\mathbf{M}^{*'}\boldsymbol{\Omega}^{-1}\mathbf{y}^{*}\\ \boldsymbol{\Sigma}_{\beta,u} &= (\mathbf{M}^{*'}\boldsymbol{\Omega}^{-1}\mathbf{M}^{*}+\mathbf{D})^{-1}\end{aligned}$$

with $\mathbf{M}^{*}=[\mathbf{W}^{*} \quad \mathbf{Z}^{*}]$ and $\mathbf{D}=\text{blockdiag}(\mathbf{0},\mathbf{D}_{u}^{-1}\otimes\mathbf{I}_{\kappa})$, again see Ruppert et al. (2003), chapter 16 for details. As before using the blockdiagonal structure of $\boldsymbol{\Omega}$ and $\mathbf{D}$ the mean and covariance above are easily computed.

In order to arrive at the complete conditional for σ_k^2, $k=1,\ldots,G$ we note that the likelihood in lemma $(3.2.1)$ is proportional to

$$\begin{aligned}&\Big(\prod_{i=1}^{n}\prod_{j=1}^{m_i}|\boldsymbol{\Sigma}_j|^{-1/2}\Big)e^{-\frac{1}{2}\sum_{i,j}(\mathbf{y}_i^{*}(t_{ij})-\boldsymbol{\mu}_i^{*}(t_{ij}))'\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_i^{*}(t_{ij})-\boldsymbol{\mu}_i^{*}(t_{ij}))}\\ \propto\ &(\sigma_k^2)^{-m/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\left(y_{ik}^{*}(t_{i1})-\mu_{ik}^{*}(t_{i1})\right)^2}{\sigma_k^2(1-\phi_k^2)^{-1}}+\sum_{j=2}^{m_i}\frac{\left(y_{ik}^{*}(t_{ij})-\mu_{ik}^{*}(t_{ij})\right)^2}{\sigma_k^2}\right)\right)\end{aligned}$$

where $\boldsymbol{\mu}_i^{*}(t_{ij}):=\mathbf{W}_i^{*}(t_{ij})\boldsymbol{\beta}+\mathbf{Z}_i^{*}(t_{ij})\mathbf{u}$ and $\mu_{ik}^{*}(t_{ij})$ denotes the k-th component of the vector. This result is derived in the proof of lemma $(3.2.1)$, see $(A.14)$. Combining this with the inverse gamma prior we arrive at the complete conditional of σ_k^2, $k=1,\ldots,G$:

$$\begin{aligned}\sigma_k^2|\mathbf{y},\mathbf{s},\boldsymbol{\Theta}\backslash\{\sigma_k^2\} \ \sim\ \text{IG}\Bigg(&A_k+\frac{m}{2},B_k+\frac{1}{2}\sum_{i=1}^{n}\Bigg(\frac{\left(y_{ik}^{*}(t_{i1})-\mu_{ik}^{*}(t_{i1})\right)^2}{(1-\phi_k^2)^{-1}}\\ &+\sum_{j=2}^{m_i}\left(y_{ik}^{*}(t_{ij})-\mu_{ik}^{*}(t_{ij})\right)^2\Bigg)\Bigg)\end{aligned}$$

The complete conditional for $\sigma_{u\ell}^2$ is identical to the one in the independent case so the following is stated without going into details:

$$\sigma_{u\ell}^2|\mathbf{y},\mathbf{s},\boldsymbol{\Theta}\backslash\{\sigma_{u\ell}^2\} \ \sim\ \text{IG}\left(A_u+\frac{\kappa}{2},B_u+\frac{1}{2}\|\mathbf{u}_\ell\|^2\right). \tag{3.23}$$

Now let us derive the complete conditional of ϕ_k, $k=1,\ldots,G+1$. The key behind the derivation is the fact that given $\mathbf{y}$, $(\boldsymbol{\beta},\mathbf{u})'$ and $(\sigma_k^2)_{1\le k\le G+1}$ the errors in

(3.12) become degenerate. Let $\varepsilon_{ik}(t_{ij}) = Y_{ik}(t_{ij}) - \mathbf{W}_{ik}(t_{ij})\boldsymbol{\beta} - \mathbf{Z}_{ik}(t_{ij})\mathbf{u}$ and focus on the likelihood based on this transformation through the model in (3.13). Let $\boldsymbol{\varepsilon}_{ik} := (\varepsilon_{ik}(t_{i1}), \ldots, \varepsilon_{ik}(t_{im_i}))'$, $\boldsymbol{\varepsilon}_{\cdot k} := (\boldsymbol{\varepsilon}_{1k}, \ldots, \boldsymbol{\varepsilon}_{nk})'$ and define

$$\begin{aligned} g(\phi_k) &= \left(\frac{\sigma_k^2}{1-\phi_k^2}\right)^{-n/2} e^{-\frac{(1-\phi_k^2)}{2\sigma_k^2}\sum_i \varepsilon_{ik}^2(t_{i1})} \\ \mu_{\phi_k} &= \frac{\sum_{i=1}^n \sum_{j=2}^{m_i} \varepsilon_{ik}(t_{ij})\varepsilon_{ik}(t_{i,j-1})}{\sum_{i=1}^n \sum_{j=2}^{m_i} \varepsilon_{ik}^2(t_{i,j-1})} \\ \sigma_{\phi_k}^2 &= \frac{\sigma_k^2}{\sum_{i=1}^n \sum_{j=2}^{m_i} \varepsilon_{ik}^2(t_{i,j-1})} \end{aligned} \tag{3.24}$$

Then by the law of total probability we can factor the likelihood associated with ϕ_k into

$$\begin{aligned} f(\boldsymbol{\varepsilon}_{\cdot k}) &= \prod_{i=1}^n \left(f(\varepsilon_{ik}(t_{i1})) \prod_{j=2}^{m_i} f(\varepsilon_{ik}(t_{ij}) | \varepsilon_{ik}(t_{i,j-1})) \right) \\ &\propto \prod_{i=1}^n \left(\left(\frac{\sigma_k^2}{1-\phi_k^2}\right)^{-1/2} e^{-\frac{(1-\phi_k^2)}{2\sigma_k^2}\varepsilon_{ik}^2(t_{i1})} e^{-\frac{1}{2\sigma_k^2}\sum_{j=2}^{m_i}(\varepsilon_{ik}(t_{ij}) - \phi_k\varepsilon_{ik}(t_{i,j-1}))^2} \right) \\ &= g(\phi_k) e^{-\frac{1}{2\sigma_k^2}\sum_{i=1}^n \sum_{j=2}^{m_i}(\phi_k^2\varepsilon_{ik}^2(t_{i,j-1}) - 2\phi_k\varepsilon_{ik}(t_{ij})\varepsilon_{ik}(t_{i,j-1}) + \varepsilon_{ik}^2(t_{ij}))} \\ &\propto g(\phi_k) e^{-\frac{1}{2\sigma_{\phi_k}^2}\left(\phi_k^2 - 2\phi_k\mu_{\phi_k} + \mu_{\phi_k}^2\right)} \end{aligned}$$

Exploring the above formula for the likelihood and combining with the prior of ϕ_k it is not hard to see that the complete conditional becomes

$$\phi_k | \mathbf{y}, \mathbf{s}, \boldsymbol{\Theta} \backslash \{\phi_k\} \sim g(\phi_k) \times \mathrm{N}(\mu_{\phi_k}, \sigma_{\phi_k}^2)\mathrm{I}_{(-1,1)}(\phi_k) \tag{3.25}$$

In order to simulate ϕ_k we can apply a Metropolis-Hastings step in the Gibbs sampling scheme. A natural candidate density would be $\mathrm{N}(\mu_{\phi_k}, \sigma_{\phi_k}^2)\mathrm{I}_{(-1,1)}(\phi_k)$, that of a truncated normal. At the r-th iteration we accept a new draw $\phi_k^{(r)}$ from the candidate density with probability $\min(g(\phi_k^{(r)})/g(\phi_k^{(r-1)}), 1)$.

3.2.3 Imputing the latent variable

Recall that the $G+1$ endogenous variables, $Y_{i,G+1}(t_{ij})$, are not observed at any time t_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$. As in the independent case we need to impute these unobserved latent responses at each iteration step of the Gibbs sampling scheme. In the independent case we could sample these values independently but in the autoregressive case we need to sample them sequentially. In the proof of lemma $(3.2.1)$ in the appendix we derive the following distributional properties

$$\begin{aligned} & \mathbf{y}_i(t_{i1})|\boldsymbol{\Theta} \\ \sim \; & \mathrm{N}\left((\mathbf{I}-\boldsymbol{\Gamma}(t_{i1}))^{-1}\boldsymbol{\Delta}(t_{i1})\mathbf{X}_i(t_{ij}), (\mathbf{I}-\boldsymbol{\Gamma}(t_{i1}))^{-1}\boldsymbol{\Sigma}_1(\mathbf{I}-\boldsymbol{\Gamma}(t_{i1}))^{-T}\right) \end{aligned} \tag{3.26}$$

and

$$\begin{aligned} & \mathbf{y}_i(t_{ij})|\mathbf{y}_i(t_{i,j-1}), \boldsymbol{\Theta} \\ \sim \; & \mathrm{N}\Bigg((\mathbf{I}-\boldsymbol{\Gamma}(t_{ij}))^{-1}\Big(\boldsymbol{\Delta}(t_{ij})\mathbf{X}_i(t_{ij}) + \boldsymbol{\Phi}\{\mathbf{Y}_i(t_{i,j-1}) - \boldsymbol{\Gamma}(t_{i,j-1})\mathbf{Y}_i(t_{i,j-1}) \\ & -\boldsymbol{\Delta}(t_{i,j-1})\mathbf{X}_i(t_{i,j-1})\}\Big), (\mathbf{I}-\boldsymbol{\Gamma}(t_{ij}))^{-1}\boldsymbol{\Sigma}_j(\mathbf{I}-\boldsymbol{\Gamma}(t_{ij}))^{-T}\Bigg) \end{aligned} \tag{3.27}$$

for all $j \geq 2$. In order to find the complete conditional of $\mathbf{y}_{G+1}$ we need to focus on the parts on the right hand side of the posterior given in (3.20) that involve $\mathbf{y}_{G+1}$. It follows from lemma $(3.2.1)$ and lemma $(3.2.2)$ that

$$\begin{aligned} \left[\mathbf{y}_{G+1}|\mathbf{y}_{obs}, \mathbf{s}, \boldsymbol{\Theta}\right] \; \propto \; & \prod_{i=1}^{n}\prod_{j=1}^{m_i} \mathrm{I}(y_{i,G+1}(t_{ij}) \in A_{ij}(s_i(t_{ij})) \\ & \times \prod_{i=1}^{n}\Big(\left[\mathbf{y}_i(t_{i1})|\boldsymbol{\Theta}\right]\prod_{j=2}^{m_i}\left[\mathbf{y}_i(t_{ij})|\mathbf{y}_i(t_{i,j-1}), \boldsymbol{\Theta}\right]\Big) \end{aligned} \tag{3.28}$$

which shows that in order to simulate from the complete conditional of $\mathbf{y}_{G+1}$ we can independently for each i start by simulating $y_{i,G+1}(t_{i1})$ from the truncated conditional normal density $\left[y_{i,G+1}(t_{i1})|\mathbf{y}_{i,obs}(t_{i1}), \boldsymbol{\Theta}\right]\mathrm{I}(y_{i,G+1}(t_{i1}) \in A_{i1}(s_i(t_{i1}))$ which is readily calculated from (3.26). Then recursively given $y_{i,G+1}(t_{i,j-1})$

we proceed to simulate $y_{i,G+1}(t_{ij})$ from the truncated conditional normal density $\big[y_{i,G+1}(t_{ij})|\mathbf{y}_{i,\text{obs}}(t_{ij}), y_{i,G+1}(t_{i,j-1}), \boldsymbol{\Theta}\big]\mathrm{I}(y_{i,G+1}(t_{ij}) \in A_{ij}(s_i(t_{ij}))$ which can be calculated from (3.27).

3.2.4 Imputing missing covariates

Assume as in the independent case that we have missing covariates $y_{i1}(t_{ij}), \ldots, y_{iG}(t_{ij})$ at time t_{ij}. The vector $\mathbf{y}_i(t_{ij}) = (y_{i1}(t_{ij}), \ldots, y_{iG}(t_{ij}), y_{i,G+1}(t_{ij}))$ is thus completely unobserved and needs to be imputed at each step of the Gibbs sampler. After imputing the latent variables $y_{i,G+1}(t_{i1}), \ldots, y_{i,G+1}(t_{i,j-1})$ recursively it follows from (3.28) that we can impute $\mathbf{y}_i(t_{ij})$ directly based on the truncated normal $\big[\mathbf{y}_i(t_{ij})|\mathbf{y}_i(t_{i,j-1}), \boldsymbol{\Theta}\big]\mathrm{I}(y_{i,G+1}(t_{ij}) \in A_{ij}(s_i(t_{ij})))$.

3.3 Simultaneous credible bands

In this section we will derive approximate simultaneous credible bands for the time-dependent parameters of (2.4):

$$\eta_k(t) = \mathbf{B}(t)\boldsymbol{\beta}_k + \mathbf{C}(t)\mathbf{u}_k, \quad k = 1, \ldots, g+d$$

More specifically, we want to construct $g+d$ $(1-\alpha)\%$ simultaneous credible bands for the $\eta_k(t)$'s such that with probability $1-\alpha$, under the posterior distribution, each curve $\eta_k(t)$ is completely contained in its corresponding credible band. We propose using a similar technique used for constructing simultaneous confidence bands in Ruppert et al. (2003). To handle all $g+d$ time-dependent parameters at once we focus on the vector $\boldsymbol{\eta}(t) = (\eta_1(t), \ldots, \eta_{g+d}(t))'$ and recall that

$$\boldsymbol{\eta}(t) \;=\; (\mathbf{I}_{g+d} \otimes \mathbf{B}(t))\boldsymbol{\beta} + (\mathbf{I}_{g+d} \otimes \mathbf{C}(t))\mathbf{u}$$

Define a grid of time points $t_1, \ldots, t_M$ at which we evaluate the lower and upper endpoints of the credible bands. Let $\boldsymbol{\beta}_1^*, \mathbf{u}_1^*, \ldots, \boldsymbol{\beta}_N^*, \mathbf{u}_N^*$ denote the output of the MCMC algorithm and define $\boldsymbol{\eta} := (\boldsymbol{\eta}(t_1)', \ldots, \boldsymbol{\eta}(t_M)')'$. Let $\boldsymbol{\eta}_1^*, \ldots, \boldsymbol{\eta}_N^*$ be the sample obtained by plugging the posterior sample of $(\boldsymbol{\beta}, \mathbf{u})$ into (3.29) at $t_1, \ldots, t_M$. It is clear that this represents a sample from the posterior distribution of $\boldsymbol{\eta}$. Now obtain an estimate of the posterior mean and standard deviation of $\eta_k(t_j)$, $\hat{\mathrm{E}}\left[\eta_k(t_j)|\mathbf{y}\right]$, and $\widehat{\text{st.dev}}\left[\eta_k(t_j)|\mathbf{y}\right]$, for all $k = 1, \ldots, g + d$ and $j = 1, \ldots, M$. We can obtain these by calculating Rao-Blackwell estimates of the posterior mean and covariance matrix of $(\boldsymbol{\beta}, \mathbf{u})$ where we recall that

$$\boldsymbol{\beta}, \mathbf{u}|\mathbf{y}, \mathbf{s}, \boldsymbol{\Theta}\backslash\{\boldsymbol{\beta}, \mathbf{u}\} \quad \sim \quad \mathbf{N}(\boldsymbol{\mu}_{\beta,u}, \boldsymbol{\Sigma}_{\beta,u})$$

and then simply evaluate (3.29) on the pre-specified grid. The standard deviations are the square roots of the diagonal elements of the resulting covariance matrix. In the following argument we ignore the variability in the estimates $\hat{\mathrm{E}}\left[\eta_k(t_j)|\mathbf{y}\right]$, and $\widehat{\text{st.dev}}\left[\eta_k(t_j)|\mathbf{y}\right]$ and consider them to approximately correspond to the true posterior mean and standard deviation. We define the random variable

$$\sup_{k,t} \left| \frac{\eta_k(t) - \mathrm{E}\left(\eta_k(t)|\mathbf{y}\right)}{\text{st.dev}\left(\eta_k(t)|\mathbf{y}\right)} \right| \approx \max_{k,j} \left| \frac{\eta_k(t_j) - \hat{\mathrm{E}}\left(\eta_k(t_j)|\mathbf{y}\right)}{\widehat{\text{st.dev}}\left(\eta_k(t_j)|\mathbf{y}\right)} \right| \tag{3.29}$$

and note that by evaluating the expression on the right for the N posterior sample points $\boldsymbol{\eta}_1^*, \ldots, \boldsymbol{\eta}_N^*$ we have an approximate sample from the posterior distribution of the expression in (3.29). Based on this sample we calculate the $(1 - \alpha)$ sample quantile and call it $m_{1-\alpha}$. It follows by construction that for all k, j the following intervals will constitute the $(1 - \alpha)\%$ simultaneous credible bands:

$$\hat{\mathrm{E}}\left[\eta_k(t_j)|\mathbf{y}\right] \pm m_{1-\alpha}\widehat{\text{st.dev}}\left[\eta_k(t_j)|\mathbf{y}\right].$$

CHAPTER 4

EXAMPLE

4.1 Head and Neck Cancer Study

We wish to analyze data arising from the head-and-neck-cancer study described in more detail in Efron (1988). There were 96 patients that entered the study and they were assigned to one of two groups, A and B. The patients of group A ($n_A = 51$) were treated with radiation therapy while the patients in group B ($n_B = 45$) got that same treatment plus chemotherapy. The Northern California Oncology Group conducted the study. The survival times in days were recorded for each patient some of which were censored. Censoring occurred mainly because patients entered the study at different calendar times and so at the end of the study some of them were still alive. The data is displayed with the Kaplan-Meier curve in Figure 4.1.

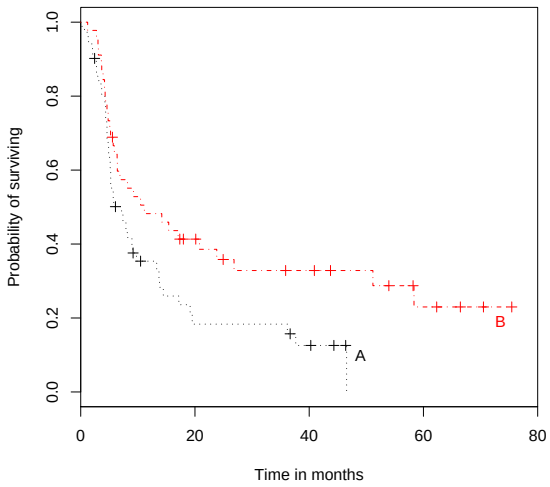

Figure 4.1: Kaplan-Meier survival estimates for arm A and B of the head-and-neck-cancer study. Discretization was based on the assumption that one month is 30.438 days and "$+$" indicates censoring.

The setup is very simple in this example as the system of equations reduces to a single probit model for the hazard. Let us start by analyzing separately each arm of the study using our methods. The model becomes

$$\Phi^{-1}(h_j) = \mu(t^*_{j-1}) \tag{4.1}$$

where $h_j = h_{ij}$ is the joint hazard for all individuals i at month j. We discretize the time-axis for arm A into monthly intervals, $[t_0, t_1) = [0, 1), \ldots, [46, 47) = [t_{m-1}, t_m)$, where one month is assumed to be 30.438 days and we let t^*_{j-1} be the midpoints of the intervals rather than the left endpoints. To compare with the results of Efron (1988) we model the intercept function with the same cubic linear spline basis

$$\mu(t^*_{j-1}) = z_j\alpha \tag{4.2}$$

where $z_j = (1, t^*_{j-1}, (t^*_{j-1} - 11)^2_-, (t^*_{j-1} - 11)^2_-)$ with $(t^*_{j-1} - 11)_- = \min(0, t^*_{j-1} - 11)$. This allows the probit-hazard to behave like a cubic polynomial in the first 11 months after which it is constrained to be linear. An important point already made is that this special case is equivalent to the approach taken in Efron (1988) if we replace the logit link used in that paper with the probit link used here. We fit the model assuming a uniform improper prior on α and run a single 3000 iteration chain in our Gibbs sampler. The resulting fit is displayed in Figure 4.2 along with a fit using the logistic regression approach of Efron (1988). We also use a probit link for the logistic regression and the fitting is done using the glm function of R. As we can see the two estimated curves go almost hand in hand at all time points, which is to be expected.

4.2 Offsets for arm B analysis

For arm B we again follow the discretization of Efron (1988) and discretize the time-axis into $1/2$ month intervals for the first 9 months, $[t_0, t_1) = [0, 1/2), \ldots,$

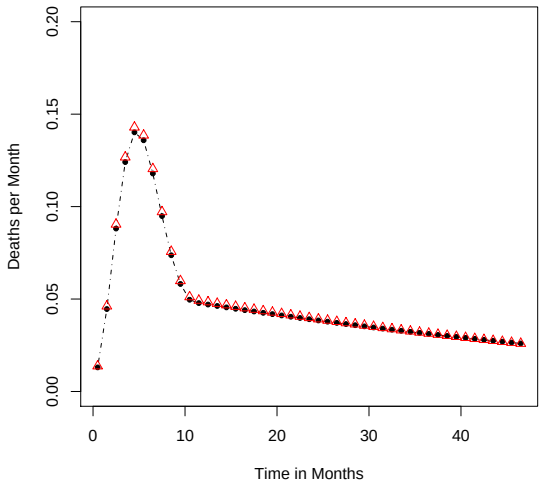

Figure 4.2: The estimated hazard function for arm A using the cubic linear spline model (4.2) fitted with the Gibbs sampler presented in this paper (smooth bullet curve) compared with the logistic regression curve of Efron (1988) with a probit link (red triangles).

$[8\ 1/2, 9) = [t_{17}, t_{18})$, then into monthly intervals until month 27, $[t_{19}, t_{20}) = [9, 10), \ldots, [26, 27) = [t_{35}, t_{36})$, and finally into 2 month intervals, $[t_{37}, t_{38}) = [27, 29), \ldots, [75, 77) = [t_{60}, t_{61})$. We model the joint hazard for all individuals according to (4.1) and (4.2) as we did for arm A. Note that in order to estimate the continuous hazard we use the approximation $h_j = h(t_{j-1})\Delta_j$, where Δ_j is the length of the jth interval. This does not affect the analysis of arm A since all interval lengths are equal to 1 but for arm B we need to be more careful. In order to account for the differing lengths of the time intervals and obtain a hazard curve comparable to the curve for arm A in Figure 4.2 we need to include the offset term discussed in section 2.2 of chapter 2. Efron (1988) considered the offset term $a_j = \log \Delta_j$, or $a_j = \log(1/2)$, $j = 1, \ldots, 18$, $a_j = \log(1)$, $j = 19, \ldots, 36$ and $a_j = \log(2)$, $j = 37, \ldots, 61$ which works under the logit-link and the approximation $\log \frac{p}{1-p} = \log p$ for small p. Note that the model

$$\log \frac{h_j}{1 - h_j} = a_j + z_j \alpha \tag{4.3}$$

implies

$$\begin{aligned}
\log \frac{h(t_{j-1})}{1-h(t_{j-1})} &= \log h(t_{j-1}) + e_{1j} \\
&= \log h_j - a_j + e_{1j} \\
&= \log \frac{h_j}{1-h_j} - a_j + e_{1j} + e_{2j} \\
&= z_j\alpha + e_j
\end{aligned}$$

where $e_j = e_{1j} + e_{2j}$ denotes error due to the approximation. Thus fitting the model (4.3) leads to the approximate estimate:

$$\hat{h}(t_{j-1}) = \frac{1}{1+\exp(-z_j\hat{\alpha})}$$

In order to fit our probit model we include the offset term $a_j = .6 \log \Delta_j$ and end up with the following model

$$\Phi^{-1}(h_j) = a_j + z_j\alpha \tag{4.4}$$

After obtaining estimates of α we get the following approximate estimate of the continuous hazard

$$\hat{h}(t_{j-1}) = \Phi(z_j\hat{\alpha})$$

Placing a uniform improper prior on α in (4.4) leads to the fit shown in Figure 4.3 and the estimate based on the logistic regression model in (4.3) is displayed for comparison.

4.2.1 Linear spline approximation of the offset

Recall that when we include the offset in our probit model we are using two approximations, one involves approximating the probit with a logit and the second involves approximating the logit with a $\log$. As an alternative consider approximating Φ^{-1} with

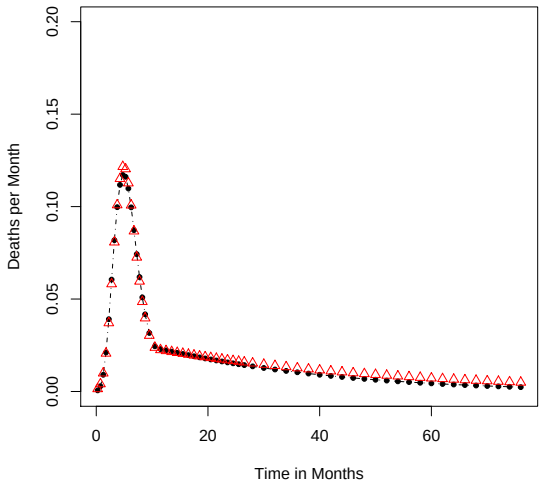

Figure 4.3: The estimated hazard function for arm B using the model (4.4) and the offset term $.5 \log \Delta_j$ (bullet curve) compared with the logistic regression model in (4.3) with offset term $\log \Delta_j$ (red triangles).

a linear spline

$$\Phi^{-1}(p) = \beta_0 + \beta_1 \log p + c_1(\log p - \tau_1)_+ \cdots + c_\kappa(\log p - \tau_\kappa)_+ + \varepsilon \tag{4.5}$$

When we fit the model

$$\Phi^{-1}(h_j) = a_j + z_j\alpha \tag{4.6}$$

for some offset term a_j, we get the following

$$\begin{aligned}
\Phi^{-1}(h(t_{j-1})) &= \beta_0 + \beta_1 \log h(t_{j-1}) + \sum_{k=1}^{\kappa} c_k(\log h(t_{j-1}) - \tau_k)_+ + e_{j1} \\
&= \beta_0 + \beta_1 \log h_j - \beta_1 \log \Delta_j \\
&\quad + \sum_{k=1}^{\kappa} c_k \mathrm{I}(\log h(t_{j-1}) > \tau_k)(\log h_j - \log \Delta_j - \tau_k) + e_{j1} \\
&= \beta_0 + \beta_1 \log h_j + \sum_{k=1}^{\kappa} c_k \mathrm{I}(\log h(t_{j-1}) > \tau_k)(\log h_j - \tau_k) \\
&\quad - \Big(\beta_1 + \sum_{k=1}^{\kappa} c_k \mathrm{I}(\log h(t_{j-1}) > \tau_k)\Big) \log \Delta_j + e_{j1} \\
&= \Phi^{-1}(h_j) + e_{2j} - a_j + e_{1j} \qquad (4.7) \\
&= z_j\alpha + e_j
\end{aligned}$$

where we let $a_j := \Big(\beta_1 + \sum_{k=1}^{\kappa} c_k \mathrm{I}(\log h(t_{j-1}) > \tau_k)\Big) \log \Delta_j$ be our offset term and $e_j = e_{1j} + e_{2j}$ is the error due to this spline approximation. Note that the approximation in (4.7) does not arise from the linear spline formula in (4.5) since $\mathrm{I}(\log h(t_{j-1}) > \tau_k) \neq \mathrm{I}(\log h_j > \tau_k)$. This makes the selection of the linear spline offset somewhat ad hoc, but leads nonetheless to reasonable results. It turns out that if we carefully select the inner knots, this approximation outperforms the logit/$\log$ approximation in (4.3) under a certain criteria explained below. Note also that the offset term is dependent on the hazard $h(t_{j-1})$ and is thus data driven. This leads to the idea of iteratively fitting the model in (4.6) and fitting the linear spline regression model (4.5) to determine the offset. The algorithm can be described as follows:

1. Determine a starting value for the offset term a_j, such as $.5 \log \Delta_j$.
2. Fit the model in (4.6) to obtain the estimate $\hat{h}(t_{j-1})$.
3. Calculate $\Phi^{-1}(\hat{h}(t_{j-1})$ and $\log \hat{h}(t_{j-1})$ for each $j = 1, \ldots, m$. Regress $\Phi^{-1}(\hat{h}(t_{j-1})$ on $\log \hat{h}(t_{j-1})$ based on (4.5) to obtain the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{c}_1, \ldots, \hat{c}_\kappa$. Set
$$a_j = \Big(\hat{\beta}_1 + \sum_{k=1}^{\kappa} \hat{c}_k \mathrm{I}(\log \hat{h}(t_{j-1}) > \tau_k)\Big) \log \Delta_j$$
4. Repeat steps 2. and 3. until the norm between two iterates of the offset vectors
$$\|\boldsymbol{a}^{(r)} - \boldsymbol{a}^{(r-1)}\| = \Big(\sum_j \Big| a_j^{(r)} - a_j^{(r-1)} \Big|^2 \Big)^{1/2}$$
becomes reasonably small.

Once we have decided on an offset term a_j, we can go ahead and fit the probit model to obtain our desired estimate $\hat{h}(t_{j-1})$. The result obtained by fitting (4.6) using a uniform improper prior on α is displayed and compared to the estimate based on

(4.3) in Figure (4.4). The offset was found by applying the above algorithm with a starting value of $a_j = .5 \log \Delta_j$. At step 3. of each iteration we selected the inner knots, τ_k, as the midpoints between $\log \hat{h}(t_{j-1})^{(k-1)}$ and $\log \hat{h}(t_{j-1})^{(k)}$ for all $k = 3, \ldots, m-1$, where $\log \hat{h}(t_{j-1})^{(k)}$ denotes the k-th component of the ordered vector $(\log \hat{h}(t_0), \ldots, \log \hat{h}(t_{m-1}))'$. The final set of knots used for the offset are the ones obtained at the last iteration of the algorithm. The algorithm converged pretty much immediately with norm difference $\|\boldsymbol{a}^{(r)} - \boldsymbol{a}^{(r-1)}\| < 10^{-8}$ for $r = 12$. In order to determine how reasonable our approximation is, we calculate

$$\begin{aligned}
\hat{e}_{1j} &= \Phi^{-1}(\hat{h}(t_{j-1})) - \left(\hat{\beta}_0 + \hat{\beta}_1 \log \hat{h}(t_{j-1}) + \sum_{k=1}^{\kappa} \hat{c}_k (\log h(t_{j-1}) - \tau_k)_+\right) \\
\hat{e}_{2j} &= \left(\hat{\beta}_0 + \hat{\beta}_1 \log \hat{h}_j + \sum_{k=1}^{\kappa} \hat{c}_k \mathbf{I}(\log \hat{h}(t_{j-1}) > \tau_k)(\log \hat{h}_j - \tau_k)\right) - \Phi^{-1}(\hat{h}_j)
\end{aligned}$$

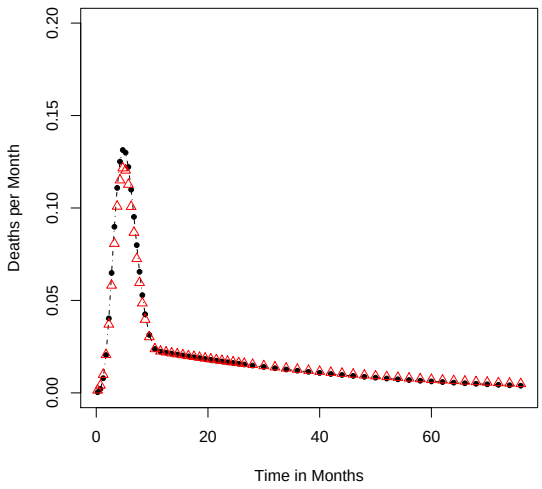

Figure 4.4: The estimated hazard function for arm B using the model (4.4) and the offset term $a_j = \left(\hat{\beta}_1 + \sum_{k=1}^{\kappa} \hat{c}_k \mathbf{I}(\log \hat{h}(t_{j-1}) > \tau_k)\right) \log \Delta_j$ (bullet curve) compared with the logistic regression model in (4.3) with offset term $\log \Delta_j$ (red triangles).

and let $\hat{e}_j = \hat{e}_{1j} + \hat{e}_{2j}$. In order to compare to the logistic regression approximation we let $\hat{e}^*_j = \hat{e}^*_{1j} + \hat{e}^*_{2j}$, where

$$\begin{aligned} \hat{e}^*_{1j} &= \log \frac{\hat{h}^*(t_{j-1})}{1-\hat{h}^*(t_{j-1})} - \log \hat{h}^*(t_{j-1}) \\ \hat{e}^*_{2j} &= \log \hat{h}^*_j - \log \frac{\hat{h}^*_j}{1-\hat{h}^*_j} \end{aligned}$$

and $\hat{h}^*(t_{j-1})$ is obtained from fitting (4.3). We get $\max_j |\hat{e}_j| \simeq 0.039$ and $\sum_j |\hat{e}_j| \simeq 0.836$ compared to $\max_j |\hat{e}^*_j| \simeq 0.067$ and $\sum_j |\hat{e}^*_j| \simeq 0.872$. We were able to reduce our error $\sum_j |\hat{e}^*_j|$ a little bit by carefully choosing the knots although that did not change the estimates much. Overall, it seemed like the linear spline approximation worked better for hazard values around the peak in Figure (4.4), whereas the logistic approximation worked better in the tales for small values of h.

4.2.2 Joint analysis

Now let us consider a joint analysis of arm A and B. By looking at Figure 4.1. we see that there is no data available for group A after the 47th month. This imposes a problem as we need to extrapolate beyond that month for treatment A. We will, however, for demonstration purposes go ahead with this analysis. The data for each patient i comprises of d_{ij} equal to 1 if the patient died in interval j, 0 otherwise and the time independent treatment indicator $X_i = 0, 1$ for arm A and B respectively. We consider the following model:

$$\Phi^{-1}(h_{ij}) = \mu_0(t^*_{j-1}) + \delta_{10}(t^*_{j-1})X_i \tag{4.8}$$

where h_{ij} is the hazard for individual i in month j and as before we have chosen t^*_{j-1} to be the midpoints of the intervals. We will partition the time-axis into one month intervals $[t_0, t_1) = [0, 1), \ldots, [75, 76) = [t_{75}, t_{76})$. The functional parameter $\mu_0(t)$ represents the

mean probit-hazard of arm A and $\delta_{10}(t)$ the difference between the mean probit-hazards of arm A and B. In the joint analysis we model the time dependent parameters with a linear spline

$$\mu_0(t) = \mathbf{B}(t)\boldsymbol{\beta}_1 + \mathbf{C}(t)\mathbf{u}_1, \quad \delta_{10}(t) = \mathbf{B}(t)\boldsymbol{\beta}_2 + \mathbf{C}(t)\mathbf{u}_2$$

where $\mathbf{B}(t) = (1, t)$ and $\mathbf{C}(t) = \big[(t-1)_+, (t-3)_+, (t-4.5)_+, (t-5)_+, (t-5.5)_+, (t-8)_+, (t-11)_+\big]$. The inner knots of the truncated spline basis are chosen so as to account for the peak around month 5. We place the following priors on the parameters

$$\begin{aligned} [\boldsymbol{\beta}_k] &\equiv 1, \\ \mathbf{u}_k &\sim \mathbf{N}(\mathbf{0}, \sigma^2_{uk}\mathbf{I}_7), \\ \sigma^2_{uk} &\sim \mathbf{IG}(0.1, 0.1), \end{aligned}$$

for $k = 1, 2$. The fitted hazards $\hat{h}_A(t) \equiv \Phi(\hat{\mu}_0(t))$ and $\hat{h}_B(t) \equiv \Phi(\hat{\mu}_0(t) + \hat{\delta}_{10}(t))$ are displayed in Figure 4.5 and as we can see the fits show the same general shape as when fitted separately. However, it seems like the peak difference is slightly more exaggerated

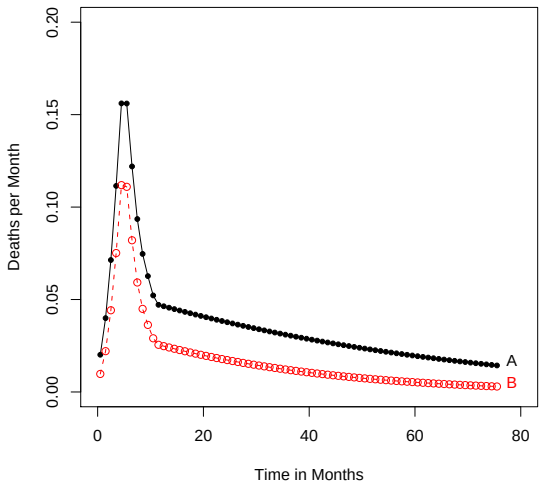

Figure 4.5: The estimated hazard functions for arm A and B from the joint model in (4.8).

from what we saw in the analysis based on separate fits. This could easily be due to the fact that here we are fitting things with a linear spline with random effects as opposed to the cubic linear spline from before.

CHAPTER 5

DISCUSSION

In this part of the thesis we have developed methods that can be applied to the estimation of time varying parameters in dynamic graphical models. We have focused on models with endogenous continuous and binary variables, that are time-dependent. Special cases include models with endogenous continuous variables only and endogenous binary variables only. Through the probit link we construct a Gaussian system of equations, which facilitates a simple estimation procedure. One of the novelties of our method is the modeling of time changing coefficients in a smooth way through splines. This leads us to a single mixed model for the system of equations whose parameters we estimate with an easily implemented Gibbs sampler. The advantage of using a Bayesian estimation procedure, rather than finding MLEs, is the simplicity of the inference procedures. The distributions of estimators of indirect and total effects are highly complicated and thus any frequentist inference about these effects need to be made through for example bootstrapping methods. From the realized Markov chain, however, we get a sample from the posterior distributions of all direct, indirect, and total effects, which allows us to construct credible bands around all these functional effects in a simple manner. It would be interesting to extend our results to a system of generalized linear models. By modeling the time changing parameters with splines we would end up with a system of generalized linear mixed models. As we have discussed, the above methods also have an application in survival analysis. Another interesting direction for this research is to consider a system of equations where continuous variables are modeled through linear regression and the survival times are modeled with a Cox proportional hazard model. It is not clear how one would carry out the estimation if we model the functional parameters with splines.

Part II

Model based clustering of microarray data

CHAPTER 6

INTRODUCTION

In recent years analyses of high dimensional data arising from Genetics have become increasingly more common in the Statistics literature. The analysis of DNA microarray data is one such example and in most cases the number of measurements is much larger than the number of individuals. An important problem in the field of Cancer epigenomics is to be able to use the information captured by epigenetic modifications to identify different biological subtypes of malignancies, such as Acute myeloid leukemia (AML), based on microarray data. AML is a highly heterogeneous disease and accurate clinical classification, risk stratification and targeted therapy of this disease remain a major challenge. Figueroa et al. (2010) performed the first large-scale DNA methylation profiling study in humans, using a data set of 344 patients with AML, collected at Erasmus University Medical Center (Rotterdam) between 1990-2008. They hypothesized that DNA methylation is not randomly distributed in AML but rather organized into highly coordinated and well-defined patterns, which reflect the distinct biological subtypes of this heterogeneous malignancy. The main goal of Figueroa et al. (2010) was to identify different subtypes of AML and isolate discriminating genes. Epigenetic marks, such as DNA methylation, play a major role in regulating gene expression and thus in determining the malignant cell's behavior. Therefore, identifying aberrant epigenetic patterns in AML and predicting methylation for different subtypes is useful in the identification of distinct forms of this disease, which might respond differently to current therapies. Furthermore, the identification of specific deregulated genes and pathways can be used to design specifically targeted therapies for the different subtypes identified.

While many biological studies have carried out gene expression profiling of AML and other malignancies, these studies have proven to be incapable of capturing all the levels of biological heterogeneity of the disease. The addition of information captured by DNA methylation has helped to rescue some of the variability lost in the noise of gene expression microarrays revealing the existence of additional levels of biological complexity. This was clearly shown in the work by Figueroa et al. (2009) in which they described the existence of two distinct forms of AML with very clear differences in their clinical outcomes despite the fact that they had been identified as a unique cohort by gene expression profiling methods. In their earlier work, Figueroa et al. (2008) had previously shown that the integration of gene expression and epigenetic platforms could be used to rescue genes that were biologically relevant but had been missed by the individual analyses of either platform. This phenomenon is explained by the fact that each of these platforms has technical and analytical limitations, which makes it impossible for them to capture the totality of the biological variability. However, by combining the data from both platforms they were able to harness greater biological strength, and identify true biological variability, as confirmed by more sensitive methods. Their work clearly demonstrated that the integration of data from different methods can be harnessed to achieve a maximum amount of biological information.

In high-dimensional data clustering is often performed on a smaller subset of the variables. In fact using all variables in high-dimensional clustering analysis has proven to give misleading results. This is for example pointed out in Tadesse et al. (2005). Clustering patients and selecting discriminating variables simultaneously is a challenging task since it seems unclear how one could identify genes that discriminate between groups if we don't know the true grouping structure. Most statistical methods do however separate these two tasks and cluster the data only after a suitable subset has been chosen. An example of such practice is McLachlan et al. (2002), where the selection

of a subset involves choosing a significance threshold for the covariates. This is essentially what is done in Figueroa et al. (2010), but there the AML patients are clustered using hierarchical correlation based clustering using a subset of the most variable genes. Another method involves dimension reduction before clustering, through principal components analysis, see for example Ghosh and Chinnaiyan (2002). The problem with selecting a subset before clustering is that we might throw away variables that contribute to unveiling the grouping structure and we might include genes that contain no information about the clustering. For microarray data there is an implicit array effect, which introduces variability from subject to subject. Thus, it is reasonable to suspect that selecting the most variable genes or using principal components analysis is not the optimal solution. There is some research on the problem of simultaneously clustering and selecting variables. Friedman and Meulman (2003) proposed a hierarchical procedure that clusters objects on separate subsets of variables, but one of the problems with that approach is that it does not guide the selection of the number of clusters. Tadesse et al. (2005) deal with this issue, but rather than allowing for different subsets of variables, they seek a single subset of discriminating variables. They propose a Bayesian approach that simultaneously selects the discriminating variables and clusters the patients into K groups, where K is unknown. For the variable selection they introduce latent binary variables that take values 1 if they define a mixture distribution for the data and 0 otherwise. Kim et al. (2006) build further on the model of Tadesse et al. (2005) and propose a Bayesian paradigm based on infinite mixtures of distributions via Dirichlet process mixtures.

In this part of the thesis we propose a model based clustering strategy that attempts to deal with both the clustering and variable selection problem simultaneously. In particular we build a finite mixture model that guides the clustering. These types of models have been shown to give a principled statistical approach to practical issues that can

come up in clustering (McLachlan and Basford (1988), Banfield and Raftery (1993), Cheeseman and Stutz (1995) and Fraley and Raftery (1998)). For a thorough review of model based clustering methods, see Fraley and Raftery (2002). In chapter 7 we introduce a hierarchical model based clustering algorithm, that can be applied to a pre-selected subset of the data. The clustering criteria involves maximizing an objective function $\pi : \mathbb{P}_n \rightarrow \mathbb{R}$, where $\mathbb{P}_n$ denotes the set of all possible partitions of the patient set $[n] = \{1, \ldots, n\}$. In our setting the objective function is a partition specific likelihood of the data and the number of clusters is automatically determined through the algorithm. A similar objective function approach can for example be found in Heard et al. (2006) and Booth et al. (2008), where they consider the clustering problem within a Bayesian framework and the objective function is the marginal posterior of the partition $\mathcal{C} \in \mathbb{P}_n$. The former authors consider a hierarchical Bayesian algorithm, whereas the latter propose a stochastic search on $\mathbb{P}_n$, starting from an initial partition. The likelihood that we construct involves a specific mixture density on the variables and estimation is carried out with the EM algorithm of Dempster et al. (1977). The hierarchical algorithm that we present guides us towards a reasonable initial guess of the true grouping structure of the patients. However, in chapter 7 we also suggest a model that imposes a mixture density on the patients and estimation of parameters in that setting can be carried out through a modified classification EM (CEM) algorithm. Through the CEM algorithm we start at the initial partition from the hierarchical algorithm and converge towards a partition with a higher likelihood. The details of the classification EM algorithm can be found in McLachlan and Peel (2000). In chapter 8 we extend the model of chapter 7 to account for all variables and address the simultaneous problem of clustering and selecting discriminating variables. To that end we define a gene-importance indicator, similar to that of Tadesse et al. (2005), that takes value 1 and 0 depending on which of two mixture distributions it defines. The mixture corresponding to the value 1 is defined

by the partition $\mathcal{C}$ and represents the density of the discriminating variables, whereas the mixture corresponding to the value 0 is not dependent on $\mathcal{C}$ and thus represents the density of variables that do not discriminate between the clusters of $\mathcal{C}$. The extended model also facilitates AML classification of new patients. One of the novelties of our clustering approach is that we allow for individual specific parameters and thus account for the microarray effect in a model based manner. These parameters are assumed fixed, but in chapter 9 we discuss a more realistic model that imposes a random distribution on the parameters. The implementation of the more realistic model is infeasible, but we show that the proposed models of chapters 7 and 8 are in fact approximations of the more realistic model. In chapter 10 we analyze the Erasmus data set and discuss how through the output of our clustering algorithm we can automatically determine which genes discriminate between the different classes. Another contribution of this part of the thesis is the integrative analysis that we suggest over multiple data platforms. We show how the proposed methods can naturally be extended to account for multiple data types and in our analysis section we show how the joint analysis of methylation and expression data outperforms the single platform analyses of each data type separately. As mentioned above each data platforms has technical and analytical limitations, making it impossible to capture the totality of the biological variability using each one separately. However, by borrowing strength across platforms biologists hope to retrieve maximum amount of biological information from the analysis. It is therefore imperative that we can bring the power of integrative methods to the field of unsupervised clustering. Such integrative analyses could reveal further layers of biological complexity in these data sets.

CHAPTER 7

MODEL BASED CLUSTERING OF METHYLATION DATA

In this chapter we introduce the model used for clustering patients with AML based on their methylation profiles. In particular we construct a likelihood for any given partition of the patient set. This partition likelihood is maximized and used as a basis for clustering and classification. In section 7.1 we specify the model and provide basic notation that will be used throughout this part of the thesis. In section 7.2 we discuss the estimation of likelihood parameters and in section 7.3 we introduce the hierarchical likelihood based clustering algorithm. The partition likelihood specification imposes a restriction on the parameters and this restriction needs to be addressed when fitting using the EM algorithm. In section 7.4 we discuss this issue in detail, in particular in the context of the hierarchical clustering algorithm. We derive some asymptotic properties of the methylation predictors in section 7.5 and in section 7.6 we discuss how the model can be extended to account for multiple data types. The hierarchical clustering algorithm of section 7.3 provides us with a good guess for the true patient partition. However, we can improve upon the obtained partition by applying the so called two way Classification EM (CEM) algorithm. We introduce and describe this two way algorithm in detail in section 7.7.

7.1 Partition Likelihood

Assume we have data consisting of methylation log-ratio measurements on G genes, or DNA fragments, of n patients with AML. A more detailed description of the data set can be found in chapter 10. For the Erasmus methylation data $G = 25,626$ and $n = 344$. In what follows we will talk about genes or DNA fragments interchangeably although the latter is more accurate since not all DNA fragments correspond to unique genes. Let

y_{ij} denote the methylation response for patient $i = 1, \ldots, n$, and gene $j = 1, \ldots, G$. Higher values of y_{ij} indicate that patient i did not methylate on gene j, whereas lower values indicate methylation. Looking at a histogram of $(y_{ij})_{j=1,\ldots,G}$ for each patient i we see roughly a bimodal distribution, see Figure 10.1 for an example of such a histogram for one of the Erasmus patients. For this type of methylation data we generally expect to see this type of bimodal mixture where the left mode corresponds to methylated genes and the right mode to genes that are not methylated. For a more detailed biological description of the methylation data see chapter 10. For the genes that are somewhere in the middle it is not clear in advance whether or not they are methylated. As many of the genes behave almost identically across subjects we try to identify beforehand a subset of genes, $J_d \subset \{1, \ldots, G\}$, that discriminate well between the different subtypes of AML. This subset of discriminating genes is then used for the clustering algorithm. Figueroa et al. (2010) suggested choosing the most variable genes among the G genes and let J_d be the set of all genes with variances greater than some threshold value, δ. In Figueroa et al. (2010) several clustering results were examined based on various values of δ and finally a value of $\delta = 1$ was chosen as it gave a sensible clustering result. In what follows we will assume that the discriminating set, J_d, represents the most variable genes and put on hold the discussion on how we can select this set in a model based manner until chapter 8. Without loss of generality we assume that the set of discriminating genes is given by $J_d = \{1, \ldots, G_\delta\}$, where G_δ represents the total number of genes with variances exceeding the threshold δ.

We now state the model assumptions and construct a likelihood for any given partition of the patient set, $[n] = \{1, \ldots, n\}$. Let us assume that $\mathcal{C}$ is the partition that represents the true subtype classification of the n patients. Methylation in AML patients is not believed to be randomly distributed across the genome but rather have systematic patterns within each subtype. This hypothesis leads us to our main model assumption

that patients within a given cluster $c \in \mathcal{C}$ have identical methylation profiles. More specifically, for any given gene, j, we assume that it either methylates for all patients in cluster c, or it does not methylate for all patients in cluster c. To formalize the assumption we introduce cluster specific methylation indicators, $\mathbf{w}_c = (w_{cj})_{j \in J_d}$, that are defined as follows:

$$w_{cj} = \begin{cases} 1 & \text{if gene } j \text{ is methylated for all patients } i \text{ in cluster } c \\ 0 & \text{if gene } j \text{ is not methylated for all patients } i \text{ in cluster } c \end{cases} \tag{7.1}$$

Note that this assumption is not going to be absolutely true in reality. However, we expect to see a consistency in methylation patterns for patients that share a cluster. For patients in cluster c we only observe the continuous responses $(y_{ij})_{i \in c}$ on gene j. This only provides an indication as to whether or not gene j is methylated for patients $i \in c$, and the above indicators are not directly observed. We thus assume a priori that the indicators for cluster c are latent random variables and put the following prior on the vector $\mathbf{w}_c$:

$$f(\mathbf{w}_c) = \prod_{j \in J_d} \pi_{1c}^{w_{cj}} \pi_{0c}^{1-w_{cj}}, \quad \pi_{1c} + \pi_{0c} = 1, \tag{7.2}$$

where π_{1c} and π_{0c} denote the proportions of genes that are methylated and not methylated, respectively, in cluster c. Now define $\mathbf{y}_i = (y_{ij})_{j=1,\ldots,G_\delta}$ as the observed methylation profile of patient i. We assume that the random variables $(\mathbf{y}_i)_{i \in c}$ are conditionally independent, given the cluster specific methylation indicators, $\mathbf{w}_c$, with densities

$$f(\mathbf{y}_i | \mathbf{w}_c) = \prod_{j \in J_d} \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2)^{w_{cj}} \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2)^{1-w_{cj}}, \tag{7.3}$$

for all $i \in c$, where ϕ denotes the normal density. This means that, conditioned on the cluster specific methylation indicators, $(y_{ij})_{j,\ldots,G_\delta}$ is a random sample arising from two normal populations, $N(\mu_{1i}, \sigma_{1i}^2)$, and $N(\mu_{2i}, \sigma_{2i}^2)$. We assume $\mu_{1i} \leq \mu_{2i}$ for all i, and since lower values of y_{ij} indicate methylation, $N(\mu_{1i}, \sigma_{1i}^2)$ represents the population of methylated genes, and $N(\mu_{2i}, \sigma_{2i}^2)$ the population of non-methylated genes. Examining

the density in (7.3) closely we see that if $w_{cj} = 1$, then y_{ij} comes from the methylated population ($y_{ij} \sim N(\mu_{1i}, \sigma^2_{1i})$) for all $i \in c$. Similarly if $w_{cj} = 0$, then y_{ij} comes from the non-methylated population ($y_{ij} \sim N(\mu_{2i}, \sigma^2_{2i})$) for all $i \in c$. We know of no biological mechanism that would imply normality but this assumption appears to be consistent with the observed data. In general we simply expect to observe bimodal mixtures for each patient and there is an explicit biological assumption of two distinct populations of methylated and non-methylated genes represented by the left mode and right mode respectively. Furthermore, it is a biological assumption that a specific gene comes from the same population (methylated or non-methylated population) for all patients sharing a distinct AML subtype, $c \in \mathcal{C}$. This leads us to the distributional assumption given in 7.3. Note also that we are allowing for individual specific parameters in the likelihood. In model based clustering the density of $\mathbf{y}_i$, on the assumption that $i \in c$, usually involves cluster specific parameters only. However, in the context of our data, having individual specific parameters makes sense as measurements on each patient, i, are taken on physically different microarray chips, which could introduce variation from subject to subject. Furthermore, if we look at histograms for patients that share a cluster, in many cases it seems highly unreasonable to assume that they have identical distributions. Let $\mathbf{y}_c = (\mathbf{y}_i)_{i \in c}$ denote the observed methylation data vector for cluster c and by the conditional independence assumption we get

$$\begin{aligned} f(\mathbf{y}_c|\mathbf{w}_c) &= \prod_{i \in c} \prod_{j \in J_d} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i})^{w_{cj}} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i})^{1-w_{cj}} \\ &= \prod_{j \in J_d} \Big(\prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i}) \Big)^{w_{cj}} \Big(\prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i}) \Big)^{1-w_{cj}}. \end{aligned} \quad (7.4)$$

Note that the random variables $(\mathbf{y}_i)_{i \in c}$ are not unconditionally independent because of the assumption that patients in a given cluster share a common methylation indicator on each gene. From (7.2) and (7.4) it is clear that the joint density of $(\mathbf{y}_c, \mathbf{w}_c)$, for each

cluster c, is given by

$$f(\mathbf{y}_c, \mathbf{w}_c) = \prod_{j \in J_d} \left(\pi_{1c} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) \right)^{w_{cj}} \left(\pi_{0c} \prod_{i \in c} \phi(y_{cij} | \mu_{2i}, \sigma_{2i}^2) \right)^{1 - w_{cj}}. \quad (7.5)$$

We let $(\mathbf{y}, \mathbf{w}) = (\mathbf{y}_c, \mathbf{w}_c)_{c \in \mathcal{C}}$ and assume independence across clusters, so that

$$f(\mathbf{y}, \mathbf{w}) = \prod_{c \in \mathcal{C}} f(\mathbf{y}_c, \mathbf{w}_c).$$

Throughout this thesis we will refer to $(\mathbf{y}, \mathbf{w})$ as the complete data and $\mathbf{y}$ as the incomplete or observed data. In order to arrive at the observed data likelihood, for the data in cluster c, we need to integrate $\mathbf{w}_c$ out from (7.5), which gives us

$$L_c = \prod_{j \in J_d} \left(\pi_{1c} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) + \pi_{0c} \prod_{i \in c} \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2) \right), \quad (7.6)$$

and the observed data likelihood of the partition $\mathcal{C}$ is then simply a product of the likelihoods for each cluster,

$$L_{\mathcal{C}} = \prod_{c \in \mathcal{C}} \prod_{j \in J_d} \left(\pi_{1c} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) + \pi_{0c} \prod_{i \in c} \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2) \right). \quad (7.7)$$

The above model construction is valid for the true partition when we can assume identical methylation within clusters and the objective of clustering is to search for this true partition, $\mathcal{C}$. In sections 7.3 and 7.7 we describe a strategy that involves searching for the partition $\mathcal{C}$ with the highest maximized likelihood $L_{\mathcal{C}}$. The maximization of the above likelihood is carried out with the EM algorithm and the procedure is detailed in the following section.

Before describing the maximization of the likelihood let us emphasize that $L_{\mathcal{C}}$ is not identifiable if we relax the constraint $\mu_{1i} \leq \mu_{2i}$, all i. This problem is well known in the mixture model literature and is called the label switching problem. It turns out that by switching the parameters π_{1c} and π_{0c} and swapping the parameter vectors $(\mu_{1i}, \sigma_{1i}^2)_{i \in c}$ and $(\mu_{2i}, \sigma_{2i}^2)_{i \in c}$ we obtain the same likelihood value. This label switching problem is

not a huge issue as we can always redefine the ordering after we reach a solution, since we know that the methylated genes have a smaller mean than the non-methylated genes. However, in the clustering context we could potentially get contradictory results if we do not put the restriction $\mu_{1i} \leq \mu_{2i}$, all i, when fitting the EM algorithm. A detailed discussion about this issue is provided in section 7.4.

7.2 EM Algorithm

The likelihood given in (7.7) does not have closed form maximizers. In fact, when we allow for unequal variances $\sigma_{1i}^2 \neq \sigma_{2i}^2$, the likelihood is unbounded and does not have a global maxima. This can be seen by setting one of the means equal to one of the data points, say $\mu_{1i} = y_{ij}$, some i, j. Then the likelihood approaches infinity as $\sigma_{1i}^2 \to 0+$. However, since the likelihood is a product of mixture distributions, derived from the distribution of the complete data, we can apply the EM algorithm of Dempster et al. (1977). This algorithm is guaranteed to achieve a local maxima if it converges. But it is important to monitor the results of the algorithm to see if in fact it did converge rather than approach a spurious solution of the type just mentioned. A nice coverage of how the EM algorithm works can for example be found in McLachlan and Peel (2000). They recommend running the EM algorithm from several different starting values, dismiss any spurious solutions, and pick the parameter values that lead to the largest likelihood value. Let $\boldsymbol{\theta}_c = ((\pi_{1c}), (\mu_{1i}, \sigma_{1i}^2, \mu_{2i}, \sigma_{2i}^2)_{i \in c})$ denote the cluster specific parameters and let $\boldsymbol{\theta} = (\boldsymbol{\theta}_c)_{c \in \mathcal{C}}$. From (7.6) and (7.7) we see that the likelihood separates, with respect to the cluster specific parameters, $L_{\mathcal{C}}(\boldsymbol{\theta}) = \prod_{c \in \mathcal{C}} L_c(\boldsymbol{\theta}_c)$. Thus, maximizing $L_{\mathcal{C}}$ simply involves maximizing L_c for each cluster c separately. In this section we will describe the steps of the EM algorithm for maximizing $L_c(\boldsymbol{\theta}_c)$ for a given cluster c. From (7.5) we can see that the complete data loglikelihood of cluster c is given by (recall $J_d =$

$\{1, \ldots, G_\delta\}$)

$$\begin{aligned} \log L_c(\boldsymbol{\theta}_c) &= \sum_{j=1}^{G_\delta} w_{cj} \Big\{ \log \pi_{1c} + \sum_{i \in c} \log \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) \Big\} \\ &\quad + (1 - w_{cj}) \Big\{ \log \pi_{0c} + \sum_{i \in c} \log \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2) \Big\}. \end{aligned} \tag{7.8}$$

The EM algorithm involves treating the $(w_{cj})_j$ as missing data and iterating between two steps, E (for expectation) and M (for maximization). Let $\boldsymbol{\theta}_c^{(0)}$ be the value specified initially for $\boldsymbol{\theta}_c$. After iterating t times between the E and the M step we arrive at the parameter iterate $\boldsymbol{\theta}_c^{(t)}$. In the next two subsections we derive the E-step and the M-step of the EM algorithm.

7.2.1 E-step

The E-step after t iterations involves taking the expectation of the complete data loglikelihood in (7.8) with respect to the density $f(\mathbf{w}_c | \mathbf{y}_c, \boldsymbol{\theta}_c^{(t)})$. But since the complete data loglikelihood is linear in the methylation indicators the E-step simply involves replacing w_{cj} in (7.8) with $\tau_{cj}^{(t)} = E[w_{cj} | \mathbf{y}_c, \boldsymbol{\theta}_c^{(t)}]$ for each j. The posterior density of $\mathbf{w}_c$ is easily derived from (7.5) and (7.6). We get

$$\begin{aligned} f(\mathbf{w}_c | \mathbf{y}_c, \boldsymbol{\theta}_c) &= \frac{\prod_{j=1}^{G_\delta} \Big(\pi_{1c} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) \Big)^{w_{cj}} \Big(\pi_{0c} \prod_{i \in c} \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2) \Big)^{1 - w_{cj}}}{\prod_{j=1}^{G_\delta} \Big(\pi_{1c} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) + \pi_{0c} \prod_{i \in c} \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2) \Big)} \\ &= \prod_{j=1}^{G_\delta} \left(\frac{\pi_{1c} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2)}{\pi_{1c} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) + \pi_{0c} \prod_{i \in c} \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2)} \right)^{w_{cj}} \\ &\quad \times \left(\frac{\pi_{0c} \prod_{i \in c} \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2)}{\pi_{1c} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) + \pi_{0c} \prod_{i \in c} \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2)} \right)^{1 - w_{cj}}, \end{aligned}$$

which is that of independent Bernoullis. It follows that the posterior expectation of w_{cj}, at a current iterate, $\boldsymbol{\theta}_c^{(t)}$, is given by

$$\tau_{cj}^{(t)} = \frac{\pi_{1c}^{(t)} \prod_{i \in c} \phi\big(y_{ij} | \mu_{1i}^{(t)}, \sigma_{1i}^{2(t)}\big)}{\pi_{1c}^{(t)} \prod_{i \in c} \phi\big(y_{ij} | \mu_{1i}^{(t)}, \sigma_{1i}^{2(t)}\big) + \pi_{0c}^{(t)} \prod_{i \in c} \phi\big(y_{ij} | \mu_{2i}^{(t)}, \sigma_{2i}^{2(t)}\big)}. \tag{7.9}$$

By plugging $\tau_{cj}^{(t)}$ into the complete data loglikelihood we arrive at the so called Q-function, which we subscript with c to emphasize that we are working with the data from cluster c,

$$\begin{aligned} Q_c(\boldsymbol{\theta}_c | \boldsymbol{\theta}_c^{(t)}) &= \sum_{j=1}^{G_\delta} \tau_{cj}^{(t)} \Big\{ \log \pi_{1c} + \sum_{i \in c} \log \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) \Big\} \\ &\quad + \big(1 - \tau_{cj}^{(t)}\big) \Big\{ \log \pi_{0c} + \sum_{i \in c} \log \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2) \Big\}. \end{aligned} \tag{7.10}$$

7.2.2 M-step

The M-step involves maximizing the Q_c-function, given in (7.10), with respect to $\boldsymbol{\theta}_c$. That turns out to be a straightforward task and closed form maximizers exist. More detailed derivation is provided in appendix B.1. By differentiating the Q_c-function with respect to π_{1c} (recall $\pi_{0c} = 1 - \pi_{1c}$) and setting to zero we get the update

$$\pi_{1c}^{(t+1)} = \frac{1}{G_\delta} \sum_{j=1}^{G_\delta} \tau_{cj}^{(t)}. \tag{7.11}$$

Differentiating the Q_c-function with respect to μ_{1i} and μ_{2i} and setting to zero leads to

$$\mu_{1i}^{(t+1)} = \frac{\sum_{j=1}^{G_\delta} \tau_{cj}^{(t)} y_{ij}}{\sum_{j=1}^{G_\delta} \tau_{cj}^{(t)}}, \tag{7.12}$$

$$\mu_{2i}^{(t+1)} = \frac{\sum_{j=1}^{G_\delta} (1 - \tau_{cj}^{(t)}) y_{ij}}{\sum_{j=1}^{G_\delta} (1 - \tau_{cj}^{(t)})}, \tag{7.13}$$

for all $i \in c$. Finally, by differentiating with respect to σ_{1i}^2 and σ_{2i}^2 and setting to zero we get for all $i \in c$

$$\sigma_{1i}^{2(t+1)} = \frac{\sum_{j=1}^{G_\delta} \tau_{cj}^{(t)} \left(y_{ij} - \mu_{1i}^{(t+1)}\right)^2}{\sum_{j=1}^{G_\delta} \tau_{cj}^{(t)}}, \tag{7.14}$$

$$\sigma_{2i}^{2(t+1)} = \frac{\sum_{j=1}^{G_\delta} \left(1 - \tau_{cj}^{(t)}\right) \left(y_{ij} - \mu_{2i}^{(t+1)}\right)^2}{\sum_{j=1}^{G_\delta} \left(1 - \tau_{cj}^{(t)}\right)}. \tag{7.15}$$

7.2.3 Implementation

The implementation of the EM algorithm involves the following stepwise procedure

1. Set $t = 0$, select initial values for the parameters, $\boldsymbol{\theta}_c^{(0)}$, and set the tolerance level at some $\varepsilon > 0$.
2. Calculate $\tau_{cj}^{(t)}$ based on (7.9), for all $j = 1, \ldots, G_\delta$.
3. Calculate $\boldsymbol{\theta}_c^{(t+1)}$ using the formulas in (7.11)-(7.15).
4. If $\|\boldsymbol{\theta}_c^{(t+1)} - \boldsymbol{\theta}_c^{(t)}\| < \varepsilon$, terminate algorithm (Alternatively one could monitor the change in the Q_c-function). Otherwise set $t = t + 1$ and move to step 2.

Good starting values can be obtained by fitting a mixture of two normals separately to each patient profile, $(y_{ij})_{j=1,\ldots,G_\delta}$, $i = 1, \ldots, n$. This results in the estimates $(\hat{\pi}_{1i}, \hat{\mu}_{1i}, \hat{\sigma}_{1i}^2, \hat{\mu}_{2i}, \hat{\sigma}_{2i}^2)_{i=1,\ldots,n}$, where we make sure that the order of the means is $\hat{\mu}_{1i} \leq \hat{\mu}_{2i}$, for each i. We let the initial values of the means and variances be

$$(\mu_{1i}^{(0)}, \sigma_{1i}^{2(0)}, \mu_{2i}^{(0)}, \sigma_{2i}^{2(0)}) = (\hat{\mu}_{1i}, \hat{\sigma}_{1i}^2, \hat{\mu}_{2i}, \hat{\sigma}_{2i}^2),$$

for each $i = 1, \ldots, n$, and let

$$\pi_{1c}^{(0)} = \frac{1}{n_c} \sum_{i \in c} \pi_{1i},$$

where n_c is the number of patients in cluster c. We have found that selecting good starting values for the means and variances is crucial for finding solutions that give a good fit. If we select the above starting values, the final MLE estimates of the means and variances are usually not too far from the starting values and give a pleasing fit, see for example the histograms in Figure 10.2. There is one subtlety, which we briefly discussed above, that needs to be addressed in the above EM algorithm. Since we are assuming that $\mu_{1i} \leq \mu_{2i}$ for all i we need to run a restricted EM algorithm. These subtleties will be addressed in section 7.4, but there we argue in favor of running the EM algorithm without any restrictions. In short, any partition of the patients that leads to convergence from the starting values above to a solution with $\mu_{1i} > \mu_{2i}$ we simply dismiss as an unreasonable candidate partition.

7.3 Hierarchical Clustering Algorithm

The goal of clustering is to find the partition that best conforms with the data under some criteria. In section 7.1 we constructed a likelihood, $L_{\mathcal{C}}$, for any given partition, $\mathcal{C}$, of the patient set $\{1, \ldots, n\}$. This leads us to our clustering criteria, which involves finding the partition that gives the highest maximized likelihood, $L_{\mathcal{C}}$. In theory we can look at all possible partitions, calculate the maximized likelihood for each one, and then pick the partition with the highest value. However, in practice that is an impossible task since the size of the space of partitions, when n is moderately large, is enormous. The number of partitions of the set $\{1, \ldots, 100\}$ has for example 116 digits. We thus propose a simple hierarchical algorithm that starts with the partition where each patient represents his/her own cluster. We calculate the likelihood for this partition and then merge the

two patients/clusters that leads to the highest value of $L_{\mathcal{C}}$. We continue merging clusters under this maximum likelihood criteria until we are left with one big cluster. Among the n partitions, that are obtained at the n merging steps, we pick the partition that has the highest value of $L_{\mathcal{C}}$. Note that this method automatically determines the number of clusters. Heard et al. (2006) used a similar approach, but they constructed a hierarchical Bayesian clustering algorithm that seeks the clustering leading to the maximum marginal posterior probability.

We will now describe the steps of algorithm in more detail to get a better feel for the computational burden involved.

1. We start with the partition $\mathcal{C}_1 = \big\{\{1\}, \{2\}, \ldots, \{n\}\big\}$. For each $\{i\} \in \mathcal{C}_1$ we run the EM algorithm as described in section 7.2 and obtain MLEs, $(\hat{\boldsymbol{\theta}}_{\{i\}})_{\{i\}\in\mathcal{C}_1}$. For each patient, i, we calculate the maximized loglikelihood

$$\hat{\ell}_{\{i\}} = \log \prod_{j \in J_d} \left(\hat{\pi}_{1i}\phi(y_{ij}|\hat{\mu}_{1i}, \hat{\sigma}^2_{1i}) + \hat{\pi}_{2i}\phi(y_{ij}|\hat{\mu}_{2i}, \hat{\sigma}^2_{2i})\right).$$

   and by summing over the patients, we obtain the maximized loglikelihood for $\mathcal{C}_1$

$$\hat{\ell}_{\mathcal{C}_1} = \sum_{i=1}^{n} \hat{\ell}_{\{i\}}.$$

   Note that this step involves running n separate EM algorithms, one for each patient.

2. The second step involves looking at all pairs of patients, $i_1, i_2 \in \{1, \ldots, n\}$, and finding the pair (i_1, i_2) for merger that leads to the highest value of $\ell_{\mathcal{C}_2}$ among partitions, $\mathcal{C}_2$, that have one cluster with two patients and the rest of the clusters are singletons:

$$\mathcal{C}_2 = \big\{\{i_1, i_2\}\big\} \cup \big\{\{s\}\big| s \neq i_1, i_2\big\}.$$

Note that such partitions have loglikelihood

$$\ell_{\mathcal{C}_2} = \ell_{\{i_1,i_2\}} + \sum_{i \neq i_1,i_2} \ell_{\{i\}},$$

and maximizing $\ell_{\mathcal{C}_2}$ involves maximizing the loglikelihood $\ell_{\{i_1,i_2\}}$ and the $n-2$ loglikelihoods $\ell_{\{i\}}$, all $i \neq i_1, i_2$. But we already maximized all the singleton loglikelihoods in step 1, so all we need to do is fit an EM algorithm to each cluster of the form $c = \{i_1, i_2\}$, $i_1, i_2 \in \{1, \ldots, n\}$. This requires fitting $\binom{n}{2}$ EM algorithms and we pick the pair (i_1, i_2) that gives the highest value of

$$\hat{\ell}_{\{i_1,i_2\}} + \sum_{i \neq i_1,i_2} \hat{\ell}_{\{i\}} = \hat{\ell}_{\mathcal{C}_1} - (\hat{\ell}_{\{i_1\}} + \hat{\ell}_{\{i_2\}}) + \hat{\ell}_{\{i_1,i_2\}},$$

where

$$\hat{\ell}_{\{i_1,i_2\}} = \log \prod_{j \in J_d} \left(\hat{\pi}_{1,\{i_1,i_2\}} \prod_{i \in \{i_1,i_2\}} \phi(y_{ij}|\hat{\mu}_{1i}, \hat{\sigma}^2_{1i}) + \hat{\pi}_{2,\{i_1,i_2\}} \prod_{i \in \{i_1,i_2\}} \phi(y_{ij}|\hat{\mu}_{2i}, \hat{\sigma}^2_{2i}) \right).$$

Call the new partition $\mathcal{C}_2$. Denote the elements of this partition with $c_1, \ldots, c_{n-1}$ and let c_{n-1} denote the recently merged cluster, $c_{n-1} = \{i_1, i_2\}$. Then clearly

$$\hat{\ell}_{\mathcal{C}_2} = \sum_{k=1}^{n-1} \hat{\ell}_{c_k}$$

3. In the third step we need to pick the pair of clusters, $c, c' \in \{c_1, \ldots, c_{n-1}\}$, that maximize

$$\hat{\ell}_{\mathcal{C}_2} - (\hat{\ell}_{\{c\}} + \hat{\ell}_{\{c'\}}) + \hat{\ell}_{\{c,c'\}},$$

but from the previous steps we already know the values of $\hat{\ell}_{\{c\}}$, $\hat{\ell}_{\{c'\}}$, for all $c, c' \in \{c_1, \ldots, c_{n-1}\}$, and $\hat{\ell}_{\{c,c'\}}$, for all pairs $c, c' \neq c_{n-1}$. So the third step only requires calculating $\hat{\ell}_{\{c,c'\}}$ for the pairs, $(c_1, c_{n-1}), \ldots, (c_{n-2}, c_{n-1})$, which means we need to run $n-2$ EM algorithms. We pick the pair of clusters to merge and form the partition

$$\mathcal{C}_3 = \{c, c'\} \cup \{s \in \{c_1, \ldots, c_{n-1}\} | s \neq c, c'\}$$

with

$$\hat{\ell}_{\mathcal{C}_3} = \hat{\ell}_{\mathcal{C}_2} - (\hat{\ell}_{\{c\}} + \hat{\ell}_{\{c'\}}) + \hat{\ell}_{\{c,c'\}}.$$

Call the elements of this partition $c_1, \ldots, c_{n-2}$ and let c_{n-2} denote the recently merged cluster.

4. We continue merging clusters in this way. The next step involves fitting $n-3$ EM Algorithms, the one after that $n-4$, and so on until we are left with one big cluster, $\mathcal{C}_n$, containing all n patients. After the algorithm has finished running we pick the partition $\mathcal{C}_K$ corresponding to the highest value of $\hat{\ell}_{\mathcal{C}_K}$.

Note that the above algorithm requires running a total of

$$n + \binom{n}{2} + (n-2) + (n-3) + \cdots + 1 \sim O(n^2)$$

EM algorithms. However, the complexity of the algorithms increases after each merger.

7.4 Restricted parameter space

One of the model assumptions is that if a gene methylates (does not methylate) for one patient in a given cluster, c, it also methylates (does not methylate) for all the other patients in that cluster. In the context of our model this is equivalent to saying $y_{ij} \sim N(\mu_{1i}, \sigma^2_{1i})$, for all $i \in c$, if $w_{cj} = 1$ and $y_{ij} \sim N(\mu_{2i}, \sigma^2_{2i})$, for all $i \in c$, if $w_{cj} = 0$. By assuming that the w_{cj} are independent Bernoullis with success probability, $\pi_{1c} = P(w_{cj} = 1)$, we arrived at the mixture likelihood (for a given partition $\mathcal{C}$), given in (7.7), and the idea of clustering involves finding the partition with the highest maximized $L_{\mathcal{C}}$. As we have discussed briefly the above likelihood is not identifiable, but more importantly in the clustering context we can get contradictory results if we relax the restriction $\mu_{1i} \leq \mu_{2i}$, for all i. In this section we will discuss a simple example

where we could reach an incorrect conclusion if this restriction is not considered when applying the EM algorithm. We will also describe how to implement the EM algorithm on a restricted parameter space and finally argue that the best strategy is to run the unconstrained version of the EM algorithm and simply dismiss any partition that gives a solution outside of the restricted parameter space.

Consider a simple example where there are only two patients, $i = 1, 2$. Assume for sake of clarity that we know the true methylation status of each gene for both patients and that in reality when any gene j is methylated for patient 1 it is not methylated for patient 2 and vice versa. Let

$$w_{ij} = \begin{cases} 1 & \text{if gene } j \text{ methylates for patient } i \\ 0 & \text{otherwise.} \end{cases}$$

for $i = 1, 2$. Then the above assumption implies $w_{1j} = 1 - w_{2j}$. Under our model assumption of identical methylation within clusters these two patients clearly represent two distinct clusters. Let us assume that the distribution of methylated genes is the same for both patients and follows $N(\mu_1, \sigma_1^2)$. Similarly assume the distribution of non-methylated genes is $N(\mu_2, \sigma_2^2)$ for both patients and assume $\mu_1 < \mu_2$. The condition $w_{1j} = 1 - w_{2j}$ implies that one of the two holds for all j (conditioned on the true values of the methylation indicators),

$$y_{1j} \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad y_{2j} \sim N(\mu_2, \sigma_2^2)$$
$$y_{1j} \sim N(\mu_2, \sigma_2^2) \quad \text{and} \quad y_{2j} \sim N(\mu_1, \sigma_1^2)$$

Let us assume that the normal densities, $N(\mu_1, \sigma_1^2)$, and $N(\mu_2, \sigma_2^2)$, are well separated from each other in the sense that

$$\phi(y|\mu_1, \sigma_1^2)\phi(y|\mu_2, \sigma_2^2) \approx 0$$

for all $y \in \mathbb{R}$, and a realization $y \sim N(\mu_1, \sigma_1^2)$ has $\phi(y|\mu_2, \sigma_2^2) \approx 0$, and similarly a realization $y \sim N(\mu_2, \sigma_2^2)$ has $\phi(y|\mu_1, \sigma_1^2) \approx 0$ with high probability. Let π_{1i} and π_{0i}

denote the proportions of methylated and non-methylated genes, respectively, in patient $i = 1, 2$, and let's assume that $\pi_{11} = \pi_{01} = \pi_{12} = \pi_{02} = 0.5$. The likelihood under the assumption that the two patients are in separate clusters, $L_{\{1\},\{2\}}$, is a product of the two quantities below

$$L_{\{1\}} = \prod_{j=1}^{G_\delta} \left(\pi_{11}\phi(y_{1j}|\mu_{11}, \sigma_{11}^2) + \pi_{01}\phi(y_{1j}|\mu_{21}, \sigma_{21}^2)\right), \tag{7.16}$$

$$L_{\{2\}} = \prod_{j=1}^{G_\delta} \left(\pi_{12}\phi(y_{2j}|\mu_{12}, \sigma_{12}^2) + \pi_{02}\phi(y_{2j}|\mu_{22}, \sigma_{22}^2)\right), \tag{7.17}$$

In terms of our setup, the EM algorithm for maximizing $L_{\{1\},\{2\}}$ will strive towards the solution

$$\begin{aligned}
&\prod_{j=1}^{G_\delta} \left(0.5\phi(y_{1j}|\mu_1, \sigma_1^2) + 0.5\phi(y_{1j}|\mu_2, \sigma_2^2)\right) \left(0.5\phi(y_{2j}|\mu_1, \sigma_1^2) + 0.5\phi(y_{2j}|\mu_2, \sigma_2^2)\right) \\
\approx &\prod_{j=1}^{G_\delta} 0.5^2\phi(y_{1j}|\mu_1, \sigma_1^2)\phi(y_{2j}|\mu_2, \sigma_2^2) + 0.5^2\phi(y_{1j}|\mu_2, \sigma_2^2)\phi(y_{2j}|\mu_1, \sigma_1^2) \\
\leq &\prod_{j=1}^{G_\delta} 0.5\phi(y_{1j}|\mu_1, \sigma_1^2)\phi(y_{2j}|\mu_2, \sigma_2^2) + 0.5\phi(y_{1j}|\mu_2, \sigma_2^2)\phi(y_{2j}|\mu_1, \sigma_1^2),
\end{aligned}$$

but that's exactly of the same form as the likelihood under the assumption that the two patients are in the same cluster, given by

$$L_{\{1,2\}} = \prod_{j=1}^{G_\delta} \left(\pi_1 \prod_{i=1}^{2} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) + \pi_0 \prod_{i=1}^{2} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)\right), \tag{7.18}$$

without the restriction $\mu_{1i} < \mu_{2i}$. If we do not put any such restrictions on the means then running the EM algorithm for maximizing $L_{\{1,2\}}$ will strive towards the solution (7.18) which is larger than the maximized $L_{\{1\},\{2\}}$. Thus we would incorrectly conclude that the two patients should be joined in one cluster rather than be in their own separate ones.

Now, consider running the EM algorithm of section 7.2 on cluster c. We suggest two ways of dealing with the restriction $\mu_{1i} \leq \mu_{2i}$, for all i. The first method involves

running the EM algorithm on the restricted parameter space

$$\boldsymbol{\Theta}_{0c} = \{(\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2) \in \mathbb{R}^2 \times \mathbb{R}_+^2 | \mu_{1i} \leq \mu_{2i}, i \in c\} \times \{\pi_{1c} \in [0,1]\},$$

that is to start with an initial value in $\boldsymbol{\Theta}_{0c}$ and make sure a new iterate never escapes the restricted parameters space. We will argue that the further restriction $\sigma_{1i}^2 = \sigma_{2i}^2$, for all $i \in c$, would ease computation. Assume we have just updated our parameters to $\boldsymbol{\theta}_c^{(n)} \in \boldsymbol{\Theta}_{0c}$. The E-step does not change and remains as stated in the previous section. Now in the M-step the Q_c-function in (7.10) is maximized for each cluster at the values in $(7.11) - (7.15)$. But note that these are maxima on the unrestricted parameter space,

$$\boldsymbol{\Theta}_c = \{(\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2) \in \mathbb{R}^2 \times \mathbb{R}_+^2 | i \in c\} \times \{\pi_{1c} \in [0,1]\},$$

and so if these maxima lead to $\mu_{1i} > \mu_{2i}$, for some $i \in c$, we will have moved out of the restricted parameter space, $\boldsymbol{\Theta}_{0c}$. Now, fix some $i \in c$ where the global maximum of the Q_c-function is attained at $\mu_{1i}^{(n+1)} > \mu_{2i}^{(n+1)}$. We then need to replace these two estimates with the constrained maximizers, which lie on the boundary of the restricted parameter space, with $\mu_{1i} = \mu_{2i}$. Note that the Q_c-function can be written as

$$\begin{aligned}
Q_c &= \sum_{j=1}^{G_\delta} \tau_{cj}^{(n)} \Big\{ \log \pi_{1c} + \sum_{i \in c} \log \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) \Big\} \\
&\quad + \big(1 - \tau_{cj}^{(n)}\big) \Big\{ \log \pi_{0c} + \sum_{i \in c} \log \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2) \Big\} \qquad (7.19) \\
&= K(\mu_{1i}, \mu_{2i}) - \frac{1}{2\sigma_{1i}^2} \sum_{j=1}^{G_\delta} \tau_{cj}^{(n)} (y_{ij} - \mu_{1i})^2 - \frac{1}{2\sigma_{2i}^2} \sum_{j=1}^{G_\delta} (1 - \tau_{cj}^{(n)}) (y_{ij} - \mu_{2i})^2 \\
&= K(\mu_{1i}, \mu_{2i}) - \frac{1}{2\sigma_{1i}^2} \sum_{j=1}^{G_\delta} \tau_{cj}^{(n)} y_{ij}^2 + \Big(\frac{1}{\sigma_{1i}^2} \sum_{j=1}^{G_\delta} \tau_{cj}^{(n)} y_{ij} \Big) \mu_{1i} - \Big(\frac{1}{2\sigma_{1i}^2} \sum_{j=1}^{G_\delta} \tau_{cj}^{(n)} \Big) \mu_{1i}^2 \\
&\quad - \frac{1}{2\sigma_{2i}^2} \sum_{j=1}^{G_\delta} (1 - \tau_{cj}^{(n)}) y_{ij}^2 + \Big(\frac{1}{\sigma_{2i}^2} \sum_{j=1}^{G_\delta} (1 - \tau_{cj}^{(n)}) y_{ij} \Big) \mu_{2i} - \Big(\frac{1}{2\sigma_{2i}^2} \sum_{j=1}^{G_\delta} (1 - \tau_{cj}^{(n)}) \Big) \mu_{2i}^2,
\end{aligned}$$

where $K(\mu_{1i}, \mu_{2i})$ denotes the remaining terms of the Q_c-function that do not depend on (μ_{1i}, μ_{2i}) for the specific fixed $i \in c$. Now, holding all other variables fixed, this represents a quadratic surface that opens downward and has a maximum value at the

values $(\mu_{1i}^{(n+1)}, \mu_{2i}^{(n+1)})$ given in (7.12) and (7.13). At the boundary of our restricted parameter space, $\mathbf{\Theta}_{0c}$, the surface defined above can be written as

$$\begin{aligned} z &= K(\mu_{1i}, \mu_{2i}) - \frac{1}{2\sigma_{1i}^2}\sum_{j=1}^{G_\delta}\tau_{cj}^{(n)}y_{ij}^2 - \frac{1}{2\sigma_{2i}^2}\sum_{j=1}^{G_\delta}(1-\tau_{cj}^{(n)})y_{ij}^2 \\ &+\Big(\frac{1}{\sigma_{1i}^2}\sum_{j=1}^{G_\delta}\tau_{cj}^{(n)}y_{ij} + \frac{1}{\sigma_{2i}^2}\sum_{j=1}^{G_\delta}(1-\tau_{cj}^{(n)})y_{ij}\Big)\mu_{1i} \\ &-\Big(\frac{1}{2\sigma_{1i}^2}\sum_{j=1}^{G_\delta}\tau_{cj}^{(n)} + \frac{1}{2\sigma_{2i}^2}\sum_{j=1}^{G_\delta}(1-\tau_{cj}^{(n)})\Big)\mu_{1i}^2, \end{aligned} \tag{7.20}$$

which after differentiating with respect to μ_{1i}, and setting to zero, leads to the maximizer

$$\mu_{1i}^{(n+1)} = \mu_{2i}^{(n+1)} = \frac{\frac{1}{2\sigma_{1i}^2}\sum_{j=1}^{G_\delta}\tau_{cj}^{(n)}y_{ij} + \frac{1}{2\sigma_{2i}^2}\sum_{j=1}^{G_\delta}(1-\tau_{cj}^{(n)})y_{ij}}{\frac{1}{2\sigma_{1i}^2}\sum_{j=1}^{G_\delta}\tau_{cj}^{(n)} + \frac{1}{2\sigma_{2i}^2}\sum_{j=1}^{G_\delta}(1-\tau_{cj}^{(n)})}$$

This maximum depends on $(\sigma_{1i}^2, \sigma_{2i}^2)$ and if we were to plug this into the Q_c-function we would require some numerical methods to maximize the function with respect to the variance components. The algorithm is computationally expensive as it is and thus it might not be feasible to do this maximization. If we however restrict the variances to be equal, $\sigma_{1i}^2 = \sigma_{2i}^2$, the solution takes on a very familiar form

$$\mu_{1i}^{(n+1)} = \mu_{2i}^{(n+1)} = \frac{1}{G_\delta}\sum_{j=1}^{G_\delta}y_{ij} = \bar{y}_{i.},$$

which leads to the joint maximizer of the variances:

$$\sigma_{1i}^{2(n+1)} = \sigma_{2i}^{2(n+1)} = \frac{1}{G_\delta}\sum_{j=1}^{G_\delta}(y_{ij} - \bar{y}_{i.})^2.$$

This solution suggests that there really is only one mode in the mixture distribution for patient i. The above method will work in general for any given partition $\mathcal{C}$, but is only feasible with the further restriction $\sigma_{1i}^2 = \sigma_{2i}^2$ for all i. This restriction on the variances does not seem reasonable and thus the restricted EM algorithm above remains an infeasible and inefficient option. We will now argue that if we use the hierarchical clustering algorithm described in section 7.3 we don't really need to worry about these

restrictions. The clustering algorithm involves first running an EM algorithm on each patient separately as if they represent their own clusters. This will result in n pairs of mean estimates (μ_{1i}, μ_{2i}), $i = 1, \ldots, n$. Since the label switching is not a problem when fitting a single mixture we simply switch labels where necessary to ensure $\mu_{1i} \leq \mu_{2i}$, for all i. In the consequent steps we merge clusters until we arrive at one cluster containing all patients. Natural starting values for $(\mu_{1i}, \mu_{2i}, \sigma^2_{1i}, \sigma^2_{2i})_i$ at each merging step are the MLEs from the previous step. We suggest running the algorithm unconstrained on each candidate cluster for merger and if the algorithm converges to a solution where $\mu_{1i} \leq \mu_{2i}$ and $\mu_{1i'} > \mu_{2i'}$ for two distinct $i, i' \in c$ we simply no longer consider that merge move an option. This will ensure that at each merging step the initial values are contained in the restricted set $\boldsymbol{\Theta}_{0c}$. Dismissing a merge that leads to a solution outside of $\boldsymbol{\Theta}_{0c}$ seems reasonable. If patients in the new merged cluster truly methylate on the same genes it seems counterintuitive that the two means of one patient would flip like that between steps. In fact our experience is that the individual mean estimates don't change much through the course of the algorithm. The same strategy can be applied when we wish to maximize the likelihood, $L_{\mathcal{C}}$, for any partition $\mathcal{C}$. We suggested in subsection 7.2.3 to initially run the EM algorithm on each patient separately and use the resulting mean and variance estimates as starting values for the EM algorithm on $\mathcal{C}$. These values in practice provide good starting values for reasonable partitions and thus we recommend running the unconstrained EM algorithm on the partition $\mathcal{C}$. If the algorithm converges to parameter estimates outside of the restricted parameter space, then we simply dismiss it as an unreasonable partition. It is worth mentioning that we have not experienced this problem of obtaining a solution that lies outside of the restricted parameter space for any of the data we have analyzed.

7.5 Asymptotics

In this section we will explore some large sample properties of the posterior expectations $(E[w_{cj}|\mathbf{y}_c, \boldsymbol{\theta}_c])_{c,j}$ of section 7.1. Note that the main asymptotic result in this section involves convergence of the posterior expectation, evaluated at the true value of the parameters $\boldsymbol{\theta}_c$, as the number of patients within cluster c goes to infinity. We have not established any asymptotic results about the estimated posterior expectation, $(E[w_{cj}|\mathbf{y}_c, \hat{\boldsymbol{\theta}}_c])_{c,j}$, as both the number of patients and number of genes go to infinity. Let's focus on a single cluster, c, and for ease of notation let n_c denote the cluster size and index the patients from $i = 1, \ldots, n_c$. Recall that $\boldsymbol{\theta}_c = ((\pi_{1c}), (\mu_{1i}, \sigma^2_{1i}, \mu_{2i}, \sigma^2_{2i})_{i \in c})$. The distributional assumptions in (7.2) and (7.4) imply that the posterior distribution of $\mathbf{w}_c$ is

$$\begin{aligned} & f(\mathbf{w}_c|\mathbf{y}_c, \boldsymbol{\theta}_c) \\ \propto \; & f(\mathbf{w}_c|\boldsymbol{\theta}_c) f(\mathbf{y}_c|\mathbf{w}_c, \boldsymbol{\theta}_c) \\ = \; & \prod_{j=1}^{G_\delta} \pi_{1c}^{w_{cj}} \pi_{0c}^{1-w_{cj}} \cdot \prod_{j=1}^{G_\delta} \Big(\prod_{i=1}^{n_c} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i}) \Big)^{w_{cj}} \Big(\prod_{i=1}^{n_c} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i}) \Big)^{1-w_{cj}}. \end{aligned}$$

By examining the above we see that as the number of patients n_c goes to infinity the information about the variables $(w_{cj})_j$ is increasing in the likelihood part but remains unchanged in the prior part. This means that in the limit as $n_c \to \infty$ the information in the likelihood part will dominate the prior information. This is formulated more rigorously in the following theorem

Theorem 7.5.1. *In section 7.2 we derived the posterior expectation of w_{cj}, for each $j = 1, \ldots, G_\delta$, (see (7.9)):*

$$E[w_{cj}|\mathbf{y}_c, \boldsymbol{\theta}_c] = \frac{\pi_{1c} \prod_{i=1}^{n_c} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i})}{\pi_{1c} \prod_{i=1}^{n_c} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i}) + \pi_{0c} \prod_{i=1}^{n_c} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i})}. \tag{7.21}$$

This is the posterior expectation of w_{cj} evaluated at the true values of the parameters and we assume that $\pi_{1c} \neq 0, 1$ and as before $\mu_{1i} < \mu_{2i}$ for all i. Let us further impose the following regularity conditions

(R.1) The variances are bounded from zero and infinity, i.e. there exist some $\lambda_1, \lambda_2 > 0$ such that $\lambda_1^2 < \sigma_{1i}^2, \sigma_{2i}^2 < \lambda_2^2$ for all i.

(R.2) The average squared mean differences are bounded away from zero and infinity, i.e. $\sup \frac{1}{n_c} \sum_{i=1}^{n_c} \Delta_i^2 < \infty$ and $\inf \frac{1}{n_c} \sum_{i=1}^{n_c} \Delta_i^2 > 0$, where $\Delta_i = \mu_{2i} - \mu_{1i} > 0$.

Under these regularity conditions

$$\lim_{n_c \to \infty} E[w_{cj} | \mathbf{y}_c, \boldsymbol{\theta}_c] = w_{cj} \quad a.s.\ P$$

where P denotes our probability measure.

Note. Regularity condition (R.1) is fairly standard but regularity condition (R.2) requires further explanation. The condition $\sup \frac{1}{n_c} \sum_{i=1}^{n_c} \Delta_i^2 < \infty$ seems very reasonable for the type of data we are working with. It simply states that the mean squared differences are bounded. The condition $\inf \frac{1}{n_c} \sum_{i=1}^{n_c} \Delta_i^2 > 0$ is necessary to prevent cases such as when the sequence Δ_i converges to 0, in which case the theorem is not true. Before proving the above theorem let us first give the following generalized version of the strong law of large numbers due to Baxter et al. (2004).

Lemma 7.5.2. *For any sequence $\{a_k\}$ with $\sup \frac{1}{n} \sum_{k=1}^{n} |a_k|^q < \infty$ for some $q > 1$*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} a_k X_k = 0 \quad a.s.$$

for every i.i.d. sequence $\{X_n\}$ with $E(|X_1|) < \infty$ and $E(X_1) = 0$.

Proof of Theorem 7.5.1. Note that the posterior expectation can be written in the following manner

$$E[w_{cj}|\mathbf{y}_c, \boldsymbol{\theta}_c] = \frac{\pi_{1c}}{\pi_{1c} + \pi_{0c} \exp\left(\sum_{i=1}^{n_c} \log \frac{\phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)}{\phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)}\right)},$$

so all we need to show is that

$$\lim_{n_c \to \infty} \sum_{i=1}^{n_c} \log \frac{\phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)}{\phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)} = \begin{cases} \infty & \text{if } w_{cj} = 0 \\ -\infty & \text{if } w_{cj} = 1 \end{cases} \quad a.s.\ P \tag{7.22}$$

(i) Conditioning on $w_{cj} = 0$, $(y_{ij})_{i=1,\ldots,n_c}$ is a sequence of independent Gaussian variables with means and variances $(\mu_{2i}, \sigma_{2i}^2)_{i=1,\ldots,n_c}$ respectively, see (7.3). We have

$$\begin{aligned}
\sum_{i=1}^{n_c} \log \frac{\phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)}{\phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)} &= \sum_{i=1}^{n_c} \left(\log \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) - \log \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)\right) \\
&= \sum_{i=1}^{n_c} \Bigg(-\frac{1}{2}\log(2\pi) - \log \sigma_{2i} - \frac{1}{2\sigma_{2i}^2}(y_{ij} - \mu_{2i})^2 \\
&\qquad + \frac{1}{2}\log(2\pi) + \log \sigma_{1i} + \frac{1}{2\sigma_{1i}^2}(y_{ij} - \mu_{1i})^2 \Bigg) \\
&= \sum_{i=1}^{n_c} \Bigg(-\frac{1}{2}\log(\sigma_{2i}^2/\sigma_{1i}^2) - \frac{1}{2\sigma_{2i}^2}(y_{ij} - \mu_{2i})^2 \\
&\qquad + \frac{1}{2\sigma_{1i}^2}(y_{ij} - \mu_{2i})^2 + \frac{1}{2\sigma_{1i}^2}(\mu_{2i} - \mu_{1i})^2 \\
&\qquad + \frac{1}{\sigma_{1i}^2}(y_{ij} - \mu_{2i})(\mu_{2i} - \mu_{1i}) \Bigg) \\
&= \frac{1}{2}\sum_{i=1}^{n_c} \left(\frac{\sigma_{2i}^2}{\sigma_{1i}^2} - 1\right)\left\{ \left(\frac{y_{ij} - \mu_{2i}}{\sigma_{2i}}\right)^2 - 1 \right\} \\
&\quad + \sum_{i=1}^{n_c} \left\{ \frac{\sigma_{2i}(\mu_{2i} - \mu_{1i})}{\sigma_{1i}^2} \right\} \left(\frac{y_{ij} - \mu_{2i}}{\sigma_{2i}}\right) \\
&\quad + \frac{1}{2}\sum_{i=1}^{n_c} \left\{ -1 + \sigma_{2i}^2/\sigma_{1i}^2 - \log(\sigma_{2i}^2/\sigma_{1i}^2) \right\} \\
&\quad + \sum_{i=1}^{n_c} \frac{1}{2\sigma_{1i}^2}(\mu_{2i} - \mu_{1i})^2
\end{aligned} \tag{7.23}$$

$$\geq \quad n_c \cdot \left\{ \frac{1}{2} \cdot \frac{1}{n_c} \sum_{i=1}^{n_c} \left(\frac{\sigma_{2i}^2}{\sigma_{1i}^2} - 1\right) \left\{ \left(\frac{y_{ij} - \mu_{2i}}{\sigma_{2i}} \right)^2 - 1 \right\} \right.$$
$$+ \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{\sigma_{2i}}{\sigma_{1i}^2} \Delta_i \left(\frac{y_{ij} - \mu_{2i}}{\sigma_{2i}} \right)$$
$$\left. + \frac{1}{2\lambda_2^2} \cdot \frac{1}{n_c} \sum_{i=1}^{n_c} \Delta_i^2 \right\}.$$

Note that the expression in (7.23) is always greater than or equal to 0. Also note that $\frac{y_{ij}-\mu_{2i}}{\sigma_{2i}}$ are i.i.d. standard normal random variables and $(\frac{y_{ij}-\mu_{2i}}{\sigma_{2i}})^2$ are i.i.d. χ_1^2 random variables. By regularity condition (R.1) $(\sigma_{2i}^2/\sigma_{1i}^2 - 1)^2$ is bounded for all i and so $\sup \frac{1}{n_c} \sum_{i=1}^{n_c} (\sigma_{2i}^2/\sigma_{1i}^2 - 1)^2 < \infty$. By regularity condition (R.1) and (R.2)

$$\sup \frac{1}{n_c} \sum_{i=1}^{n_c} \left(\frac{\sigma_{2i}}{\sigma_1^2} \Delta_i \right)^2 \leq (\lambda_2^2/\lambda_1^4) \cdot \sup \frac{1}{n_c} \sum_{i=1}^{n_c} \Delta_i^2 < \infty.$$

Hence all the conditions of lemma 7.5.2 are fulfilled and we have

$$\lim_{n_c \to \infty} \frac{1}{n_c} \sum_{i=1}^{n_c} \left(\frac{\sigma_{2i}^2}{\sigma_{1i}^2} - 1\right) \left\{ \left(\frac{y_{ij} - \mu_{2i}}{\sigma_{2i}} \right)^2 - 1 \right\} = 0 \quad a.s.\ P_0,$$

and

$$\lim_{n_c \to \infty} \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{\sigma_{2i}}{\sigma_1^2} \Delta_i \left(\frac{y_{ij} - \mu_{2i}}{\sigma_{2i}} \right) = 0 \quad a.s.\ P_0,$$

where P_0 is the conditional probability measure defined by

$$P_0(A) = P(A \cap [w_{cj} = 0])/P(w_{cj} = 0).$$

By the above and regularity condition (R.2) we now get

$$\lim_{n_c \to \infty} \sum_{i=1}^{n_c} \log \frac{\phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)}{\phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)} \geq \lim_{n_c \to \infty} n_c \cdot \left(\frac{1}{2\lambda_2^2} \cdot \inf \frac{1}{n_c} \sum_{i=1}^{n_c} \Delta_i^2 \right) = \infty \quad a.s.\ P_0$$

(ii) Conditioning on $w_{cj} = 1$, $(y_{ij})_{i=1,\ldots,n_c}$ is a sequence of independent Gaussian variables with means and variances $(\mu_{1i}, \sigma_{1i}^2)_{i=1,\ldots,n_c}$ respectively, see (7.3). We

have

$$
\begin{aligned}
\sum_{i=1}^{n_c} \log \frac{\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2)}{\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2)} &= \sum_{i=1}^{n_c} \big(\log \phi(y_{ij}|\mu_{2i},\sigma_{2i}^2) - \log \phi(y_{ij}|\mu_{1i},\sigma_{1i}^2)\big) \\
&= \sum_{i=1}^{n_c} \bigg(-\frac{1}{2}\log(2\pi) - \log\sigma_{2i} - \frac{1}{2\sigma_{2i}^2}(y_{ij}-\mu_{2i})^2 \\
&\qquad + \frac{1}{2}\log(2\pi) + \log\sigma_{1i} + \frac{1}{2\sigma_{1i}^2}(y_{ij}-\mu_{1i})^2 \bigg) \\
&= \sum_{i=1}^{n_c} \bigg(\frac{1}{2}\log(\sigma_{1i}^2/\sigma_{2i}^2) + \frac{1}{2\sigma_{1i}^2}(y_{ij}-\mu_{1i})^2 \\
&\qquad - \frac{1}{2\sigma_{2i}^2}(y_{ij}-\mu_{1i})^2 - \frac{1}{2\sigma_{2i}^2}(\mu_{1i}-\mu_{2i})^2 \\
&\qquad - \frac{1}{\sigma_{2i}^2}(y_{ij}-\mu_{1i})(\mu_{1i}-\mu_{2i}) \bigg) \\
&= \frac{1}{2}\sum_{i=1}^{n_c} \big(1-\frac{\sigma_{1i}^2}{\sigma_{2i}^2}\big)\Big\{ \left(\frac{y_{ij}-\mu_{1i}}{\sigma_{1i}}\right)^2 - 1\Big\} \\
&\quad - \sum_{i=1}^{n_c} \Big\{ \frac{\sigma_{1i}(\mu_{1i}-\mu_{2i})}{\sigma_{2i}^2} \Big\} \left(\frac{y_{ij}-\mu_{1i}}{\sigma_{1i}}\right) \\
&\quad - \frac{1}{2}\sum_{i=1}^{n_c} \big\{ -1 + \sigma_{1i}^2/\sigma_{2i}^2 - \log(\sigma_{1i}^2/\sigma_{2i}^2) \big\} \qquad (7.24) \\
&\quad - \sum_{i=1}^{n_c} \frac{1}{2\sigma_{2i}^2}(\mu_{2i}-\mu_{1i})^2 \\
&\leq n_c \cdot \bigg\{ \frac{1}{2}\cdot\frac{1}{n_c}\sum_{i=1}^{n_c} \big(1-\frac{\sigma_{1i}^2}{\sigma_{2i}^2}\big)\Big\{ \left(\frac{y_{ij}-\mu_{1i}}{\sigma_{1i}}\right)^2 - 1\Big\} \\
&\qquad + \frac{1}{n_c}\sum_{i=1}^{n_c} \frac{\sigma_{1i}}{\sigma_{2i}^2}\Delta_i \left(\frac{y_{ij}-\mu_{1i}}{\sigma_{1i}}\right) \\
&\qquad - \frac{1}{2\lambda_2^2}\cdot\frac{1}{n_c}\sum_{i=1}^{n_c}\Delta_i^2 \bigg\}.
\end{aligned}
$$

Note that the expression in (7.24) is always less than or equal to 0. Also note that $\frac{y_{ij}-\mu_{1i}}{\sigma_{1i}}$ are i.i.d. standard normal random variables and $(\frac{y_{ij}-\mu_{1i}}{\sigma_{1i}})^2$ are i.i.d. χ_1^2 random variables. By regularity condition (R.1) $(1-\sigma_{1i}^2/\sigma_{2i}^2)^2$ is bounded for all i and so $\sup \frac{1}{n_c}\sum_{i=1}^{n_c}(1-\sigma_{1i}^2/\sigma_{2i}^2)^2 < \infty$. By regularity condition (R.1) and (R.2)

$$
\sup \frac{1}{n_c}\sum_{i=1}^{n_c} \left(\frac{\sigma_{1i}}{\sigma_{2i}^2}\Delta_i\right)^2 \leq (\lambda_2^2/\lambda_1^4)\cdot \sup \frac{1}{n_c}\sum_{i=1}^{n_c}\Delta_i^2 < \infty.
$$

Hence all the conditions of lemma 7.5.2 are fulfilled and we have

$$\lim_{n_c\to\infty}\frac{1}{n_c}\sum_{i=1}^{n_c}\left(1-\frac{\sigma_{1i}^2}{\sigma_{2i}^2}\right)\left\{\left(\frac{y_{ij}-\mu_{1i}}{\sigma_{1i}}\right)^2-1\right\}=0\quad a.s.\ P_1,$$

and

$$\lim_{n_c\to\infty}\frac{1}{n_c}\sum_{i=1}^{n_c}\frac{\sigma_{1i}}{\sigma_{2i}^2}\Delta_i\left(\frac{y_{ij}-\mu_{1i}}{\sigma_{1i}}\right)=0\quad a.s.\ P_1,$$

where P_1 is the conditional probability measure defined by

$$P_1(A)=P(A\cap[w_{cj}=1])/P(w_{cj}=1).$$

By the above and regularity condition (R.2) we now get

$$\lim_{n_c\to\infty}\sum_{i=1}^{n_c}\log\frac{\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2)}{\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2)}\leq\lim_{n_c\to\infty}n_c\cdot\left(-\frac{1}{2\lambda_2^2}\cdot\inf\frac{1}{n_c}\sum_{i=1}^{n_c}\Delta_i^2\right)=-\infty\quad a.s.\ P_1$$

By (i) and (ii) above we see that (7.22) holds, which implies

$$\lim_{n_c\to\infty}E[w_{cj}|\mathbf{y}_c,\boldsymbol{\theta}_c]=\begin{cases}0 & \text{if } w_{cj}=0\\ 1 & \text{if } w_{cj}=1\end{cases}\quad a.s.\ P$$

and that proves our claim. □

7.6 Multiple platforms

Recall that the Erasmus study involved $n = 344$ patients. Methylation profiles, $(y_{ij})_{j=1,\ldots,G}$, on $G = 25,626$ different DNA fragments were measured for each patient $i = 1,\ldots,n$. In the Erasmus study, expression data was also collected on each of the $n = 344$ patients. The expression data involves measurements on $54,675$ DNA fragments and exhibits the same kind of bimodal behavior as the methylation data. In fact it turns out that assuming that patients, within a specific cluster, have identical expression profiles is also a valid biological assumption. In this section we will talk about how the

partition likelihood can easily be extended to include both data types. More generally, we discuss how we can extend the partition likelihood to account for multiple data types as long as each data type can reasonably be modeled by the model described in the previous sections. We will start by introducing the notation and then briefly summarize the above methods when extended to multiple platforms.

For patient $i = 1, \ldots, n$ we let y_{ijk} denote the response on DNA fragment $j = 1, \ldots, G_k$ in platform $k = 1, \ldots, m$. From now on we will talk about DNA fragments being either ON or OFF in each platform. In the case of two platforms, with methylation data on one and expression data on the other, we for example define ON as methylation and expression respectively. As before we let $\mathcal{C}$ denote the true partition of the n subjects. We assume patients in a given cluster $c \in \mathcal{C}$ have identical ON/OFF profiles on each platform $k = 1, \ldots, m$ independently. We thus define a cluster and platform specific indicator for each DNA fragment

$$w_{cjk} = \begin{cases} 1 & \text{if DNA fragment } j \text{ on platform } k \text{ is ON in cluster } c \\ 0 & \text{if DNA fragment } j \text{ on platform } k \text{ is OFF in cluster } c \end{cases} \tag{7.25}$$

Let J_d^k denote the set of discriminating DNA fragments for the data on platform k and without loss of generality let us assume that $J_d^k = \{1, \ldots, G_k^{\delta_k}\}$. In the context of section 7.1 we can think of $G_k^{\delta_k}$ as the number of DNA fragments from platform k that exceed the threshold δ_k in variability. Define $\mathbf{w}_c = (w_{cjk})_{j \in J_d^k, k=1,\ldots,m}$ and we assume a priori that the ON/OFF indicators for cluster c are independent Bernoullis, both across platforms and fragments,

$$f(\mathbf{w}_c) = \prod_{k=1}^{m} \prod_{j \in J_d^k} \pi_{1ck}^{w_{cjk}} \pi_{0ck}^{1-w_{cjk}}, \quad \pi_{1ck} + \pi_{0ck} = 1 \tag{7.26}$$

where π_{1ck} and π_{0ck} represent the proportions of DNA fragments on platform k that are ON and OFF, respectively, in cluster c. Now define the vector $\mathbf{y}_i = (y_{ijk})_{j \in J_d^k, k=1,\ldots,m}$

and we assume that if $i \in c$ it follows the conditional density

$$f(\mathbf{y}_i|\mathbf{w}_c) = \prod_{k=1}^{m} \prod_{j \in J_d^k} \phi(y_{ijk}|\mu_{1ik}, \sigma_{1ik}^2)^{w_{cjk}} \phi(y_{ijk}|\mu_{2ik}, \sigma_{2ik}^2)^{1-w_{cjk}}, \tag{7.27}$$

where we are assuming independence across platforms. Define $\mathbf{y}_c = (\mathbf{y}_i)_{i \in c}$ as the observed data of cluster c and we assume conditional independence across patients, so we get

$$f(\mathbf{y}_c|\mathbf{w}_c) \;=\; \prod_{k=1}^{m} \prod_{j \in J_d} \Big(\prod_{i \in c} \phi(y_{ijk}|\mu_{1ik}, \sigma_{1ik}^2) \Big)^{w_{cjk}} \Big(\prod_{i \in c} \phi(y_{ijk}|\mu_{2ik}, \sigma_{2ik}^2) \Big)^{1-w_{cjk}}.$$

The complete data likelihood of the data in cluster c is

$$f(\mathbf{y}_c, \mathbf{w}_c) = \prod_{k=1}^{m} \prod_{j \in J_d^k} \left(\pi_{1ck} \prod_{i \in c} \phi(y_{ijk}|\mu_{1ik}, \sigma_{1ik}^2) \right)^{w_{cjk}} \left(\pi_{0ck} \prod_{i \in c} \phi(y_{ijk}|\mu_{2ik}, \sigma_{2ik}^2) \right)^{1-w_{cjk}},$$

which leads the cluster specific marginal likelihood

$$L_c = \prod_{k=1}^{m} \prod_{j \in J_d^k} \left(\pi_{1ck} \prod_{i \in c} \phi(y_{ijk}|\mu_{1ik}, \sigma_{1ik}^2) + \pi_{0ck} \prod_{i \in c} \phi(y_{ijk}|\mu_{2ik}, \sigma_{2ik}^2) \right). \tag{7.28}$$

As in the the single platform case the marginal likelihood of the partition $\mathcal{C}$ is a product of the likelihoods for each cluster and if we pull the platform product outside we get

$$L_{\mathcal{C}} = \prod_{k=1}^{m} \prod_{c \in \mathcal{C}} \prod_{j \in J_d^k} \left(\pi_{1ck} \prod_{i \in c} \phi(y_{ijk}|\mu_{1ik}, \sigma_{1ik}^2) + \pi_{0ck} \prod_{i \in c} \phi(y_{ijk}|\mu_{2ik}, \sigma_{2ik}^2) \right). \tag{7.29}$$

If we compare the above likelihood to the single platform one we see that it's a product of platform specific likelihoods that all have the form given in (7.7). The independence assumption across platforms might not be valid as for each platform we have measurements on the same genome. We nonetheless assume independence due to the dimensionality of the data, which makes specification of a dependence structure across platforms infeasible for implementation. Furthermore, in cases where the number of DNA fragments, G_k, differs across platforms it is not clear how one would model the dependence in a simple way. Even if the platforms are dependent, the above likelihood

in (7.29) can be viewed as a composite likelihood (composed of likelihoods for each data type) where inference about the marginal parameters is carried out as described above. The general class of composite likelihoods was introduced by Lindsay (1988) and includes the pseudolikelihood of Besag (1974). In the literature likelihoods of the type (7.29) are sometimes called independence likelihoods (Chandler and Bate (2007)) and are for example useful in cases where dependence makes model fitting infeasible.

Maximizing $L_{\mathcal{C}}$ in the multiple platform setting simply involves fitting an EM algorithm to maximize the single platform likelihood for each platform, k, separately. The estimation procedures for maximizing a single platform likelihood are detailed in section 7.2. The hierarchical clustering algorithm in the multiple platform setting works essentially in the same way as described in section 7.3, except now the objective function to be maximized, $L_{\mathcal{C}}$, is given by (7.29) rather than (7.7). It seems that by using multiple data types we should be getting more information about the true grouping structure of our patients. In chapter 10 we do an analysis on both single and multiple platforms for methylation and expression data and we see that the multiple platforms analysis indeed has more discriminating power.

7.7 Two way Classification EM algorithm

In section 7.3 we described a hierarchical likelihood based algorithm for clustering the AML patients. The clustering criteria involves trying to find the partition $\mathcal{C}$ that achieves the highest maximized likelihood $\hat{L}_{\mathcal{C}}$. Clearly the hierarchical algorithm is not guaranteed to find the global maximum on the space of all possible partitions. In this section we will discuss a method that starts from an initial partition and then moves around the partition space in a systematic way in an attempt to find a better partition. For ease of

notation we only consider a single platform but the method extends naturally to multiple platforms as well. Note that by maximizing $L_{\mathcal{C}}$ through the use of the EM algorithm we also obtained estimates of the posterior expectations of the cluster specific methylation indicators, $(E[w_{cj}|\mathbf{y}])_{c,j}$. Thus, along the way we are also clustering the genes into two groups, methylated and non-methylated genes. In this section we introduce latent cluster membership indicators for the patients as well and develop an EM algorithm based on a fixed number of clusters for the patient partition. This algorithm involves assigning patients into clusters based on posterior expectations of the cluster membership indicators.

Given a specific partition $\mathcal{C}$ we define the methylation indicators, $(\mathbf{w}_c)_{c\in\mathcal{C}}$, in the same way as before,

$$w_{cj} = \begin{cases} 1 & \text{if gene } j \text{ is methylated for all patients } i \text{ in cluster } c \\ 0 & \text{if gene } j \text{ is not methylated for all patients } i \text{ in cluster } c \end{cases} \tag{7.30}$$

and then we define for each patient i and cluster c

$$X_{ic} = \begin{cases} 1 & \text{if patient } i \text{ is in cluster } c \\ 0 & \text{otherwise} \end{cases}$$

Note that for each i exactly one of the components of $\mathbf{X}_i = (X_{ic})_{c\in\mathcal{C}}$ is equal to 1 and the remaining ones are 0. Also note that these cluster membership indicators fully define the partition $\mathcal{C}$. We assume that the data arises from the following conditional density

$$\begin{aligned} f(\mathbf{y}|\mathbf{w},\mathbf{X},\boldsymbol{\theta}) &= \prod_{c\in\mathcal{C}}\prod_{i=1}^{n}\Big(f(\mathbf{y}_i|\mathbf{w}_c,\boldsymbol{\theta}_i)\Big)^{X_{ic}} \qquad (7.31)\\ &= \prod_{c\in\mathcal{C}}\prod_{j\in J_d}\Big(\prod_{i=1}^{n}\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2)^{X_{ic}}\Big)^{w_{cj}}\Big(\prod_{i=1}^{n}\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2)^{X_{ic}}\Big)^{1-w_{cj}}, \end{aligned}$$

where we assumed

$$f(\mathbf{y}_i|\mathbf{w}_c,\boldsymbol{\theta}_i) = \prod_{j\in J_d}\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2)^{w_{cj}}\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2)^{1-w_{cj}}. \tag{7.32}$$

We put the following Multinom$\{1, \mathbf{p} = (p_c)_{c\in\mathcal{C}}\}$ prior on $\mathbf{X} = (\mathbf{X}_i)_{i=1}^n$

$$f(\mathbf{X}) = \prod_{i=1}^{n}\prod_{c\in\mathcal{C}} p_c^{X_{ic}}, \quad \sum_{c\in\mathcal{C}} p_c = 1. \tag{7.33}$$

If we were to put a Bernoulli prior on the methylation indicators as before (see (7.2)) then we could arrive at the observed data likelihood by integrating out $\mathbf{X}$ and $\mathbf{w}$ and hope to proceed as before. However, this observed data likelihood is intractable and furthermore using an EM algorithm to maximize it is also infeasible. Notice that the complete data loglikelihood $\log f(\mathbf{y}, \mathbf{w}, \mathbf{X}|\boldsymbol{\theta})$ is easily derived from the above and is linear in terms of the form $X_{ic}w_{cj}$. The E-step would involve taking expectation of the complete data loglikelihood with respect to the distribution $f(\mathbf{w}, \mathbf{X}|\mathbf{y})$ but this distribution does not have a nice form and a closed form solution to the E-step does not exist. It is interesting to note though that $f(\mathbf{w}|\mathbf{X}, \mathbf{y})$ does have a nice form, that of independent Bernoullis, and so does $f(\mathbf{X}|\mathbf{w}, \mathbf{y})$, that of independent multinomials. We could thus in theory look into some Monte Carlo EM methods or possibly run a fully Bayesian Analysis. In this thesis we will not explore any Bayesian alternatives and as we've mentioned before, any extra computational burden is unappealing, and thus a Monte Carlo EM does not seem a feasible option.

To circumvent the above problem we resolve to the assumption of fixed methylation indicators, $(w_{cj})_{c,j}$, rather than random. We thus treat $\mathbf{w}$ as just another parameter and look into maximizing the likelihood

$$f(\mathbf{y}|\mathbf{w}, \boldsymbol{\theta}) = \prod_{i=1}^{n}\sum_{c\in\mathcal{C}} p_c f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\theta}_i), \tag{7.34}$$

where $f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\theta}_i)$ is given by (7.32). This likelihood has a familiar form and can easily be maximized with the EM algorithm. From (7.31) and (7.33), assuming now $\mathbf{w}$ is fixed,

we derive the complete data loglikelihood,

$$\begin{aligned}\log f(\mathbf{y}, \mathbf{X}|\mathbf{w}, \boldsymbol{\theta}) &= \sum_{c\in\mathcal{C}}\sum_{i=1}^{n} X_{ic}\log p_c \\ &\quad + \sum_{c\in\mathcal{C}}\sum_{j\in J_d} w_{cj}\sum_{i=1}^{n} X_{ic}\log\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2) \\ &\quad + \sum_{c\in\mathcal{C}}\sum_{j\in J_d}(1-w_{cj})\sum_{i=1}^{n} X_{ic}\log\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2),\end{aligned}$$

and we can now derive the E-step and the M-step of the EM algorithm.

7.7.1 E-step

From (7.31) and (7.33) it is clear that the posterior distribution of $\mathbf{X}$ is a product of multinomials

$$f(\mathbf{X}|\mathbf{y}) = \prod_{i=1}^{n}\prod_{c\in\mathcal{C}}\left(\frac{p_c f(\mathbf{y}_i|\mathbf{w}_c,\boldsymbol{\theta}_i)}{\sum_k p_k f(\mathbf{y}_i|\mathbf{w}_k,\boldsymbol{\theta}_i)}\right)^{X_{ic}}.$$

At a current iterate of the parameter estimates, $(w_{cj}^{(t)})_{c,j}$, $(\boldsymbol{\theta}_i^{(t)})_i$, and $(p_c^{(t)})_c$, we define the posterior expectation of X_{ic} as

$$\kappa_{ic}^{(t)} = \frac{p_c^{(t)} f(\mathbf{y}_i|\mathbf{w}_c^{(t)},\boldsymbol{\theta}_i^{(t)})}{\sum_k p_k^{(t)} f(\mathbf{y}_i|\mathbf{w}_k^{(t)},\boldsymbol{\theta}_i^{(t)})}. \tag{7.35}$$

Define $\boldsymbol{\Lambda} = \{(p_c)_c, (\mu_{1i},\mu_{2i},\sigma_{1i},\sigma_{2i}^2)_i, (w_{cj})_{c,j}\}$ and we arrive at the Q-function

$$\begin{aligned}Q(\boldsymbol{\Lambda}|\boldsymbol{\Lambda}^{(t)}) &= \sum_{c\in\mathcal{C}}\sum_{i=1}^{n}\kappa_{ic}^{(t)}\log p_c \\ &\quad + \sum_{c\in\mathcal{C}}\sum_{j\in J_d} w_{cj}\sum_{i=1}^{n}\kappa_{ic}^{(t)}\log\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2) \\ &\quad + \sum_{c\in\mathcal{C}}\sum_{j\in J_d}(1-w_{cj})\sum_{i=1}^{n}\kappa_{ic}^{(t)}\log\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2).\end{aligned} \tag{7.36}$$

We now need to maximize the above Q-function with respect to the parameters. The details of the maximization is given in the next subsection.

7.7.2 M-step

It is easy to verify that if we differentiate the Q-function with respect to p_c (keeping in mind that $\sum_c p_c = 1$) and set to zero we arrive at the following updating formula for the cluster membership proportions,

$$p_c^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \kappa_{ic}^{(t)}, \tag{7.37}$$

for all $c \in \mathcal{C}$. However, maximizing the Q-function with respect to $(w_{cj})_{c,j}$ and $(\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2)_i$ we need an iterative procedure. This iterative procedure is identical in nature to the so called Classification ML approach, which is described in section 2.21 of McLachlan and Peel (2000). In the general mixture model framework the classification ML approach involves maximizing the complete data likelihood, rather than the observed mixture likelihood, with respect to the unknown indicators $(w_{cj})_{c,j}$ as well as the parameters. In the following we do exactly that except we do not have any proportion parameters, $(\pi_{1c})_c$. The idea is quite simple, we iterate between the two following steps until we reach convergence and update the parameters to $(\mu_{1i}^{(t+1)}, \mu_{2i}^{(t+1)}, \sigma_{1i}^{2(t+1)}, \sigma_{2i}^{2(t+1)})_i$ and $(w_{cj}^{(t+1)})_{c,j,k}$ accordingly. McLachlan and Peel (2000) explain how this procedure leads to a local maxima if we converge to a non-spurious solution.

(a) For a current value of $(w_{cj})_{c,j}$ the Q-function is maximized at the following values of the parameters

$$\hat{\mu}_{1i} = \frac{\sum_{c\in\mathcal{C}} \kappa_{ic}^{(t)} \sum_{j\in J_d} w_{cj} y_{ij}}{\sum_{c\in\mathcal{C}} \kappa_{ic}^{(t)} \sum_{j\in J_d} w_{cj}}, \tag{7.38}$$

$$\hat{\mu}_{2i} = \frac{\sum_{c\in\mathcal{C}} \kappa_{ic}^{(t)} \sum_{j\in J_d} (1-w_{cj}) y_{ij}}{\sum_{c\in\mathcal{C}} \kappa_{ic}^{(t)} \sum_{j\in J_d} (1-w_{cj})}, \tag{7.39}$$

and

$$\hat{\sigma}_{1i}^2 = \frac{\sum_{c\in\mathcal{C}} \kappa_{ic}^{(t)} \sum_{j\in J_d} w_{cj}(y_{ij}-\hat{\mu}_{1i})^2}{\sum_{c\in\mathcal{C}} \kappa_{ic}^{(t)} \sum_{j\in J_d} w_{cj}}, \tag{7.40}$$

$$\hat{\sigma}_{2i}^2 = \frac{\sum_{c\in\mathcal{C}} \kappa_{ic}^{(t)} \sum_{j\in J_d} (1-w_{cj})(y_{ij}-\hat{\mu}_{2i})^2}{\sum_{c\in\mathcal{C}} \kappa_{ic}^{(t)} \sum_{j\in J_d} (1-w_{cj})}. \tag{7.41}$$

The above formulas are easily obtained by differentiating the Q-function with respect to these parameters and setting to zero.

(b) It is easy to see that for current values of $(\hat{\mu}_{1i}, \hat{\mu}_{2i}, \hat{\sigma}_{1i}^2, \hat{\sigma}_{2i}^2)_i$ the Q-function is maximized at $w_{cj} = 1$ or 0 according to whether

$$\sum_{i=1}^{n} \kappa_{ic}^{(t)} \log \phi(y_{ij}|\hat{\mu}_{1i}, \hat{\sigma}_{1i}^2) > \sum_{i=1}^{n} \kappa_{ic}^{(t)} \log \phi(y_{ij}|\hat{\mu}_{2i}, \hat{\sigma}_{2i}^2)$$

holds or not.

7.7.3 Implementation

Now the two way clustering procedure can be summarized as follows:

1. Set $t = 0$, start with initial values $\mathbf{\Lambda}^{(0)} = \{(p_c^{(0)})_c, (\mu_{1i}^{(0)}, \mu_{2i}^{(0)}, \sigma_{1i}^{2(0)}, \sigma_{2i}^{2(0)})_i, (w_{cj}^{(0)})_{c,j}\}$, and select a tolerance level ε.
2. Calculate $\kappa_{ic}^{(t)}$ as defined in (7.35) for each i, c.
3. Update $p_c^{(t+1)}$ according to (7.37) for all c.
4. Iterate between (a) and (b) in section 7.7.2 until convergence is reached to get the updates $(\mu_{1i}^{(t+1)}, \mu_{2i}^{(t+1)}, \sigma_{1i}^{2(t+1)}, \sigma_{2i}^{2(t+1)})_i$ and $(w_{cj}^{(t+1)})_{c,j}$.
5. If $\|\mathbf{\Lambda}^{(t+1)} - \mathbf{\Lambda}^{(t)}\| < \varepsilon$ terminate algorithm. Otherwise set $t = t + 1$ and return to step 2.

The implementation of the above algorithm is simplified by replacing step 2. above with the following modified E-step

2*. Calculate $\kappa_{ic}^{(t)}$ as defined in (7.35) for each i, c, but let $\kappa_{ic}^{(t)} = 1$ if $\kappa_{ic}^{(t)} > \kappa_{ic'}^{(t)}$ for all $c' \neq c$ and $\kappa_{ic}^{(t)} = 0$ otherwise.

For the data we have been working with, usually at each step, the vector $(\kappa_{ic}^{(t)})_{c \in \mathcal{C}}$ has one value close to or numerically equal to 1 and the other values close to 0. Thus we recommend replacing step 2 with step 2^* to facilitate easier implementation and a more efficient algorithm. Note that by doing so we are in fact applying the classification ML approach and maximizing the complete data loglikelihood $\log f(\mathbf{y}, \mathbf{X}|\mathbf{w}, \boldsymbol{\theta})$. Thus the above procedure is a CEM algorithm for determining the cluster memberships of the patients. The name "Two way CEM algorithm" arises from that fact that in the M-step of the patient CEM algorithm we are effectively running a CEM algorithm to determine the cluster membership of the genes (methylated vs. non-methylated genes).

CHAPTER 8

CLUSTERING AND VARIABLE SELECTION

In chapter 7 we introduced a model based algorithm that involved constructing a likelihood for any given partition, $\mathcal{C}$. One drawback to the method was that we could not use all the available data in the model. Instead we needed to determine a discriminating set of genes in advance and exclude all other genes in the analysis. This is unsatisfactory as we do not have a nice way of choosing this set of discriminating genes. In the analysis of Figueroa et al. (2010) all genes with variance exceeding a certain threshold δ were declared discriminating. It is not clear how the threshold δ should be chosen and furthermore there is no guarantee that any threshold will give the best solution to the problem. As we have discussed in chapter 7 the measurements on each patient are taken on physically different microarray chips which introduces variability from subject to subject. Thus, some genes might have high variability just due to the microarray effect. By selecting variables using the thresholding approach, we suspect that many of the genes declared discriminating are in fact noise and that many genes that are excluded as noise could add valuable information to the analysis. In this chapter we extend the partition likelihood model of chapter 7 to account for all genes. We introduce a latent gene importance indicator which will be predicted along with the methylation indicators. In the process we get information on which genes are driving the differences between the subtypes and which genes are simply noise. By doing this we can filter out the noise in a model based way and cluster the patients using all the data. In section 8.1 we present the gene importance indicators and the extended likelihood. We discuss the estimation of likelihood parameters in section 8.2 and give formulas for the posterior expectations of the latent indicators of the model. Some asymptotic results for the posterior expectations are given in section 8.3 and in section 8.4 we introduce an extended two way CEM clustering algorithm, similar to the one of section 7.7. In section 8.5 we show how the

likelihood model can also be applied to classification problems and finally in section 8.6 we will talk about how all the methods in this chapter can be extended to multiple data types.

8.1 Extended partition likelihood

In section 7.1 we constructed a partition likelihood for a subset of the data. More specifically, we chose a subset, J_d, of genes that we believed to discriminate well between classes and constructed a likelihood for the data J_d. In this section we will extend this likelihood to all genes and attempt to find the discriminating genes automatically through the model fitting. Let $\mathcal{C}$ denote the true partition of the patient set and let us assume there is a subset $J_d(\mathcal{C}) \subset \{1, \ldots, G\}$ of genes that are important for unveiling the true grouping structure of the patients. We give a more rigorous definition of this set below. For genes in the complimentary set, $\overline{J_d}(\mathcal{C})$, we assume they either methylate for all patients in the study or they do not methylate for all patients. The genes in $\overline{J_d}(\mathcal{C})$ are not expected to contain much information about the grouping structure of the patients and in that sense we declare them unimportant or noise. We call the genes in $J_d(\mathcal{C})$ important and define them by the following: $j \in J_d(\mathcal{C})$ if the density of $(y_{ij})_{i=1,\ldots,n}$ is given by the partition mixture density of section 7.1,

$$\prod_{c \in \mathcal{C}} \Big(\pi_{1c} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) + \pi_{0c} \prod_{i \in c} \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2) \Big). \tag{8.1}$$

We let $j \in \overline{J_d}(\mathcal{C})$ if the density of $(y_{ij})_{i=1,\ldots,n}$ is given by

$$\pi_1 \prod_{i=1}^{n} \phi(y_{ij} | \alpha_{1i}, \varsigma_{1i}^2) + \pi_0 \prod_{i=1}^{n} \phi(y_{ij} | \alpha_{2i}, \varsigma_{2i}^2), \tag{8.2}$$

which is of the same form as (8.1) if $\mathcal{C} = \{\{1, \ldots, n\}\}$. We interpret π_{1c}, for each c, as the proportion of important genes that are methylated in cluster c and π_1 as the proportion of unimportant genes that are methylated. We put the restrictions $\mu_{1i} < \mu_{2i}$

and $\alpha_{1i} < \alpha_{2i}$ for all i. The parameters μ_{1i} and α_{1i} represent the means of the important and unimportant methylation-gene populations, respectively. Similarly the parameters μ_{2i} and α_{2i} represent the means of the important and unimportant non-methylation-gene populations, respectively. It is biologically reasonable to assume that $(\mu_{1i}, \sigma^2_{1i}, \mu_{2i}, \sigma^2_{2i}) = (\alpha_{1i}, \varsigma^2_{1i}, \alpha_{2i}, \varsigma^2_{2i})$. Then we can interpret the parameters for each patient i as the global means and variances of the methylated and non-methylated gene populations of patient i. For the sake of generality we will however not make this assumption in deriving the methods below. Note that the set of discriminating genes depends on the partition $\mathcal{C}$. This is clear since the two densities in (8.1) and (8.2) are of the same form if $\mathcal{C}$ consists of only one big cluster, in which case there is no notion of important versus unimportant genes. The idea is now to assume that the observed methylation data, on all G genes, follows a mixture distribution with (8.1) and (8.2) as the mixture components. We then predict which of the two components each gene is more likely to come from.

For each gene, j, define an indicator, γ_j, that is equal to 1 if gene j discriminates between patients and 0 otherwise. More specifically, let

$$\gamma_j = \begin{cases} 1 & \text{if } j \in J_d(\mathcal{C}) \\ 0 & \text{if } j \in \overline{J_d}(\mathcal{C}) \end{cases}$$

and we note that the two sets can be defined as

$$\begin{aligned} J_d(\mathcal{C}) &= \{j|(y_{ij})_{i=1,\ldots,n} \text{ follows the density } (8.1)\}, \\ \overline{J_d}(\mathcal{C}) &= \{j|(y_{ij})_{i=1,\ldots,n} \text{ follows the density } (8.2)\}. \end{aligned}$$

Since we don't know in advance which genes discriminate between patients, and which don't, these indicators are latent and we will from now on refer to them as gene importance indicators. The idea of gene importance indicators was introduced in Tadesse et al.

(2005). Assume a priori that the gene importance indicators are independent Bernoullis,

$$f(\boldsymbol{\gamma}) = \prod_{j=1}^{G} p^{\gamma_j}(1-p)^{1-\gamma_j}, \tag{8.3}$$

where we interpret p as the proportion of important genes. Based on the above assumptions we have

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\gamma}) &= \prod_{j=1}^{G} \left\{ \prod_{c\in\mathcal{C}} \Big(\pi_{1c} \prod_{i\in c} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i}) + \pi_{0c} \prod_{i\in c} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i}) \Big) \right\}^{\gamma_j} \\ &\quad \times \left\{ \pi_1 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}, \varsigma^2_{1i}) + \pi_0 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{2i}, \varsigma^2_{2i}) \right\}^{1-\gamma_j}. \end{aligned} \tag{8.4}$$

Multiplying together the two densities in (8.3) and (8.4) and integrating out $\boldsymbol{\gamma}$ we arrive at the marginal likelihood

$$\begin{aligned} f(\mathbf{y}) &= \prod_{j=1}^{G} \left\{ p \prod_{c\in\mathcal{C}} \Big(\pi_{1c} \prod_{i\in c} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i}) + \pi_{0c} \prod_{i\in c} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i}) \Big) \right. \\ &\quad \left. + (1-p) \Big(\pi_1 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}, \varsigma^2_{1i}) + \pi_0 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{2i}, \varsigma^2_{2i}) \Big) \right\}. \end{aligned} \tag{8.5}$$

To facilitate an EM algorithm that is easily implemented we also define methylation indicators for the given partition $\mathcal{C}$,

$$w_{cj} = \begin{cases} 1 & \text{if gene } j \text{ is methylated in cluster } c \\ 0 & \text{if gene } j \text{ is not methylated in cluster } c \end{cases}$$

and put the following conditional prior on $\mathbf{w}_j = (w_{cj})_{c\in\mathcal{C}}$

$$f(\mathbf{w}_j|\gamma_j) = \Big(\prod_{c\in\mathcal{C}} \pi_{1c}^{w_{cj}} \pi_{0c}^{1-w_{cj}} \Big)^{\gamma_j} \Big(I(\mathbf{w}_j \in A) \pi_1^{w_{1j}} \pi_0^{1-w_{1j}} \Big)^{1-\gamma_j}, \tag{8.6}$$

where $A = \{\mathbf{w}_j | w_{cj} = w_{c'j}, \text{ all } c, c' \in \mathcal{C}\}$. We further assume conditional independence across all genes

$$f(\mathbf{w}|\boldsymbol{\gamma}) = \prod_{j=1}^{G} f(\mathbf{w}_j|\gamma_j).$$

Note that when $\gamma_j = 1$ we have a product of independent cluster specific Bernoullis in (8.6), $w_{cj} \sim bernoulli(\pi_{1c})$, but when $\gamma_j = 0$ we force all the cluster specific methylation indicators to be identical, which happens when there is only one big cluster. This

specifies the prior on the latent indicators w_{cj} for both the discriminating genes, $\gamma_j = 1$, and the noisy genes, $\gamma_j = 0$. To finalize the construction of the complete data likelihood we assume that the conditional distribution of $\mathbf{y}|\mathbf{w}, \boldsymbol{\gamma}$ is given by the density

$$\begin{aligned} f(\mathbf{y}|\mathbf{w}, \boldsymbol{\gamma}) &= \prod_{j=1}^{G} \left(\prod_{c \in \mathcal{C}} \Big\{ \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \Big\}^{w_{cj}} \Big\{ \prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) \Big\}^{1-w_{cj}} \right)^{\gamma_j} \\ &\quad \times \left(\Big\{ \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}, \varsigma_{1i}^2) \Big\}^{w_{1j}} \Big\{ \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{2i}, \varsigma_{2i}^2) \Big\}^{1-w_{1j}} \right)^{1-\gamma_j}, \end{aligned} \tag{8.7}$$

which along with (8.3) and (8.6) leads to the complete data likelihood

$$\begin{aligned} & f(\mathbf{y}, \mathbf{w}, \boldsymbol{\gamma}) \\ = & \prod_{j=1}^{G} \left(p \prod_{c \in \mathcal{C}} \Big\{ \pi_{1c} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \Big\}^{w_{cj}} \Big\{ \pi_{0c} \prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) \Big\}^{1-w_{cj}} \right)^{\gamma_j} \\ & \times \left((1-p) I(\mathbf{w}_j \in A) \Big\{ \pi_1 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}, \varsigma_{1i}^2) \Big\}^{w_{1j}} \Big\{ \pi_0 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}, \varsigma_{1i}^2) \Big\}^{1-w_{1j}} \right)^{1-\gamma_j}. \end{aligned}$$

In appendix B.2 we verify that the above leads to the marginal specified in (8.5).

Now let us write up the complete data log-likelihood and note the simple structure of it. It turns out that the E-step simply requires an application of the tower property of conditional expectation and the M-step has closed form maximizers just like in the EM algorithm of chapter 7. The parameters that need to be estimated are $\boldsymbol{\theta} = \{p, (\pi_{1c})_{c \in \mathcal{C}}, \pi_1, (\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2)_i, (\alpha_{1i}, \alpha_{2i}, \varsigma_{1i}^2, \varsigma_{2i}^2)_i\}$, and the complete data log-likelihood is given by

$$\begin{aligned} \log f(\mathbf{y}, \mathbf{w}, \boldsymbol{\gamma}|\boldsymbol{\theta}) &= \sum_{j=1}^{G} \Big(\gamma_j \log p + (1-\gamma_j) \log(1-p) \Big) \\ &+ \sum_{j=1}^{G} \sum_{c \in \mathcal{C}} \Big(\gamma_j w_{cj} \log \pi_{1c} + \gamma_j (1-w_{cj}) \log \pi_{0c} \Big) \\ &+ \sum_{j=1}^{G} \sum_{c \in \mathcal{C}} \Big(\gamma_j w_{cj} \sum_{i \in c} \log \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \\ &\qquad + \gamma_j (1-w_{cj}) \sum_{i \in c} \log \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) \Big) \end{aligned} \tag{8.8}$$

$$
\begin{aligned}
&+ \sum_{j=1}^{G}(1-\gamma_j)\log\big(I(w_{cj}=w_{c'j},\ \text{all } c,c'\in\mathcal{C})\big)\\
&+ \sum_{j=1}^{G}\Big((1-\gamma_j)w_{1j}\log\pi_1+(1-\gamma_j)(1-w_{1j})\log\pi_0\Big)\\
&+ \sum_{j=1}^{G}\Big((1-\gamma_j)w_{1j}\sum_{i=1}^{n}\log\phi(y_{ij}|\alpha_{1i},\varsigma_{1i}^2)\\
&\qquad\qquad +(1-\gamma_j)(1-w_{1j})\sum_{i=1}^{n}\log\phi(y_{ij}|\alpha_{2i},\varsigma_{2i}^2)\Big).
\end{aligned}
$$

Remark 8.1.1. *Note that the likelihood in (8.7) is not identifiable with respect to the parameters and the indicators. If we let* $(\mu_{1i},\sigma_{1i}^2,\mu_{2i},\sigma_{2i}^2)=(\alpha_{1i},\varsigma_{1i}^2,\alpha_{2i},\varsigma_{2i}^2)$*, then we get the same value for the* j*th component of* $f(\mathbf{y}|\mathbf{w},\boldsymbol{\gamma})$ *when* $\gamma_j=1$ *and* $w_{cj}=1$*, all* c*, as when* $\gamma_j=0$ *and* $w_{1j}=1$*. However, this is not a problem for the parameter estimation as the marginal likelihood (8.5) is identifiable with respect to the parameters. Note also that the condition* $\gamma_j=1$ *and* $w_{cj}=w_{c'j}$*, for all* c,c'*, can occur on a particular gene. At first glance it might seem counterintutive that you have a gene that is important* $(\gamma_j=1)$ *and yet does not discriminate between the classes (*$w_{cj}=w_{c'j}$*, all* c,c'*). However, recall the definition of* γ_j*. It simply states that* $\gamma_j=1$ *if* $(y_{ij})_{i=1,\ldots,n}$ *comes from density (8.1) rather than (8.2). Thus the above condition is not unreasonable as it implies that most likely gene* j *methylates differently for at least two patients* $i,i'\in\{1,\ldots,n\}$ *but does not necessarily discriminate between the classes of the partition.*

8.2 EM algorithm

We will now explain how the EM algorithm can be applied to estimate the parameters of the marginal likelihood (8.5). In the following subsections we will derive the Q-function of the E-step and the closed form solutions of the M-step. Technical details

will be provided in appendix B.2.

8.2.1 E-step

At a given iterate of the parameters, $\boldsymbol{\theta}^{(t)}$, the E-step of the EM algorithm involves taking the expectation of the complete data loglikelihood in (8.8) with respect to the density $f(\mathbf{w}, \boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\theta}^{(t)}) = f(\mathbf{w}|\boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\theta}^{(t)}) f(\boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\theta}^{(t)})$. By using the tower property of conditional expectation, we first take expectation with respect to $f(\mathbf{w}|\boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\theta}^{(t)})$ and then with respect to $f(\boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\theta}^{(t)})$. The 5th term of (8.8), the one involving the indicator function, does not have any parameters and thus does not affect the E-step or the M-step of the EM algorithm. In appendix B.2 we will argue that the term in fact vanishes in the E-step.

For the remaining terms, however, note that given the importance indicators, $\boldsymbol{\gamma}$, the posterior probability that gene j methylates in cluster c is

$$\begin{aligned} E[w_{cj}|\boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\theta}^{(t)}] &= \gamma_j \frac{\pi_{1c}^{(t)} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}^{(t)}, \sigma_{1i}^{2(t)})}{\pi_{1c}^{(t)} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}^{(t)}, \sigma_{1i}^{2(t)}) + \pi_{0c}^{(t)} \prod_{i \in c} \phi(y_{ij}|\mu_{2i}^{(t)}, \sigma_{2i}^{2(t)})} \\ &\quad + (1-\gamma_j) \frac{\pi_1^{(t)} \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}^{(t)}, \varsigma_{1i}^{2(t)})}{\pi_1^{(t)} \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}^{(t)}, \varsigma_{1i}^{2(t)}) + \pi_0^{(t)} \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{2i}^{(t)}, \varsigma_{2i}^{2(t)})} \\ &= \gamma_j \tau_{cj}^{(t)} + (1-\gamma_j) \nu_{1j}^{(t)}, \end{aligned} \tag{8.9}$$

say. This posterior probability is derived in appendix B.2. The posterior probability that gene j discriminates between patients is given by

$$\eta_j^{(t)} \equiv E[\gamma_j|\mathbf{y}, \boldsymbol{\theta}^{(t)}] = \frac{p^{(t)} f_1^{(t)}(\mathbf{y}_j)}{p^{(t)} f_1^{(t)}(\mathbf{y}_j) + (1-p^{(t)}) f_2^{(t)}(\mathbf{y}_j)}, \tag{8.10}$$

where

$$f_1^{(t)}(\mathbf{y}_j) = \prod_{c\in\mathcal{C}} \Big(\pi_{1c}^{(t)} \prod_{i\in c} \phi(y_{ij}|\mu_{1i}^{(t)}, \sigma_{1i}^{2(t)}) + \pi_{0c}^{(t)} \prod_{i\in c} \phi(y_{ij}|\mu_{2i}^{(t)}, \sigma_{2i}^{2(t)})\Big),$$

and

$$f_2^{(t)}(\mathbf{y}_j) = \pi_1^{(t)} \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}^{(t)}, \varsigma_{1i}^{2(t)}) + \pi_0^{(t)} \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{2i}^{(t)}, \varsigma_{2i}^{2(t)}).$$

The posterior probability in (8.10) is also derived in appendix B.2. When we take expectation of the complete data loglikelihood with respect to $f(\mathbf{w}|\boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\theta}^{(t)})$ we replace w_{cj} in (8.8) with the posterior expectation in (8.9) and note that $\gamma_j(1-\gamma_j) = 0$, $\gamma_j^2 = \gamma_j$, and $(1-\gamma_j)^2 = 1-\gamma_j$, for all j. The expression becomes linear in γ_j (since we are ignoring the vanishing $5th$ term). Next we take expectation with respect to $f(\boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\theta}^{(t)})$ and that simply involves replacing γ_j with (8.10). We now arrive at the Q-function:

$$\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{j=1}^{G} \Big(\eta_j^{(t)} \log p + (1-\eta_j^{(t)}) \log(1-p)\Big) \\
&+ \sum_{j=1}^{G} \sum_{c\in\mathcal{C}} \Big(\eta_j^{(t)} \tau_{cj}^{(t)} \log \pi_{1c} + \eta_j^{(t)}(1-\tau_{cj}^{(t)}) \log \pi_{0c}\Big) \\
&+ \sum_{j=1}^{G} \sum_{c\in\mathcal{C}} \Big(\eta_j^{(t)} \tau_{cj}^{(t)} \sum_{i\in c} \log \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \\
&\qquad + \eta_j^{(t)}(1-\tau_{cj}^{(t)}) \sum_{i\in c} \log \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)\Big) \\
&+ \sum_{j=1}^{G} \Big((1-\eta_j^{(t)}) \nu_{1j}^{(t)} \log \pi_1 + (1-\eta_j^{(t)})(1-\nu_{1j}^{(t)}) \log \pi_0\Big) \\
&+ \sum_{j=1}^{G} \Big((1-\eta_j^{(t)}) \nu_{1j}^{(t)} \sum_{i=1}^{n} \log \phi(y_{ij}|\alpha_{1i}, \varsigma_{1i}^2) \\
&\qquad + (1-\eta_j^{(t)})(1-\nu_{1j}^{(t)}) \sum_{i=1}^{n} \log \phi(y_{ij}|\alpha_{2i}, \varsigma_{2i}^2)\Big).
\end{aligned} \tag{8.11}$$

8.2.2 M-step

The M-step involves maximizing the Q-function in (8.11). By differentiating with respect to the parameters and setting to zero we get a closed form solution for each parameter. The updating formula for the proportion of important genes is

$$p^{(t+1)} = \frac{1}{G}\sum_{j=1}^{G}\eta_j^{(t)}. \tag{8.12}$$

The updating formula for the proportion of important genes that are methylated in cluster c is

$$\pi_{1c}^{(t+1)} = \frac{\sum_{j=1}^{G}\eta_j^{(t)}\tau_{cj}^{(t)}}{\sum_{j=1}^{G}\eta_j^{(t)}}. \tag{8.13}$$

This can be interpreted as the expected number of important genes that are methylated in cluster c over the expected number of important genes. Similarly for the unimportant genes we have

$$\pi_{1}^{(t+1)} = \frac{\sum_{j=1}^{G}(1-\eta_j^{(t)})\nu_{1j}^{(t)}}{\sum_{j=1}^{G}(1-\eta_j^{(t)})}, \tag{8.14}$$

which is interpreted as the expected number of unimportant genes that are methylated over the expected number of unimportant genes. As for the individual means and variances of the normal mixture components of the discriminating density (8.1) we have

$$\begin{aligned}
\mu_{1i}^{(t+1)} &= \frac{\sum_{j=1}^{G}\eta_j^{(t)}\tau_{cj}^{(t)}y_{ij}}{\sum_{j=1}^{G}\eta_j^{(t)}\tau_{cj}^{(t)}}, \\
\mu_{2i}^{(t+1)} &= \frac{\sum_{j=1}^{G}\eta_j^{(t)}(1-\tau_{cj}^{(t)})y_{ij}}{\sum_{j=1}^{G}\eta_j^{(t)}(1-\tau_{cj}^{(t)})},
\end{aligned} \tag{8.15}$$

and

$$\begin{aligned}
\sigma_{1i}^{2(t+1)} &= \frac{\sum_{j=1}^{G} \eta_j^{(t)} \tau_{cj}^{(t)} \left(y_{ij} - \mu_{1i}^{(t+1)}\right)^2}{\sum_{j=1}^{G} \eta_j^{(t)} \tau_{cj}^{(t)}}, \\
\sigma_{2i}^{2(t+1)} &= \frac{\sum_{j=1}^{G} \eta_j^{(t)} (1 - \tau_{cj}^{(t)}) \left(y_{ij} - \mu_{2i}^{(t+1)}\right)^2}{\sum_{j=1}^{G} \eta_j^{(t)} (1 - \tau_{cj}^{(t)})}.
\end{aligned} \tag{8.16}$$

Similarly for the means and variances of the non-discriminating density (8.2)

$$\begin{aligned}
\alpha_{1i}^{(t+1)} &= \frac{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) \nu_{1j}^{(t)} y_{ij}}{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) \nu_{1j}^{(t)}}, \\
\alpha_{2i}^{(t+1)} &= \frac{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) (1 - \nu_{1j}^{(t)}) y_{ij}}{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) (1 - \nu_{1j}^{(t)})},
\end{aligned} \tag{8.17}$$

and

$$\begin{aligned}
\varsigma_{1i}^{2(t+1)} &= \frac{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) \nu_{1j}^{(t)} \left(y_{ij} - \alpha_{1i}^{(t+1)}\right)^2}{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) \nu_{1j}^{(t)}}, \\
\varsigma_{2i}^{2(t+1)} &= \frac{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) (1 - \nu_{1j}^{(t)}) \left(y_{ij} - \alpha_{2i}^{(t+1)}\right)^2}{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) (1 - \nu_{1j}^{(t)})}.
\end{aligned} \tag{8.18}$$

8.2.3 M-step, equal means and variances

In the special case when the individual means and variances of the two mixture densities in (8.1) and (8.2) are equal, i.e. $(\mu_{1i}^*, \sigma_{1i}^{2*}, \mu_{2i}^*, \sigma_{2i}^{2*}) := (\mu_{1i}, \sigma_{1i}^2, \mu_{2i}, \sigma_{2i}^2) = (\alpha_{1i}, \varsigma_{1i}^2, \alpha_{2i}, \varsigma_{2i}^2)$ for all i, the E-step above remains the same but the M-step changes.

The updating formulas for $p^{(t+1)}$, $(\pi_{1c}^{(t+1)})_c$, and $\pi_1^{(t+1)}$ remain the same but the updating formula for the joint means and variances now become

$$\begin{aligned}
\mu_{1i}^{*(t+1)} &= \frac{\sum_{j=1}^{G}\left[\eta_j^{(t)}\tau_{cj}^{(t)} + (1-\eta_j^{(t)})\nu_{1j}^{(t)}\right]y_{ij}}{\sum_{j=1}^{G}\left[\eta_j^{(t)}\tau_{cj}^{(t)} + (1-\eta_j^{(t)})\nu_{1j}^{(t)}\right]}, \\
\mu_{2i}^{*(t+1)} &= \frac{\sum_{j=1}^{G}\left[\eta_j^{(t)}(1-\tau_{cj}^{(t)}) + (1-\eta_j^{(t)})(1-\nu_{1j}^{(t)})\right]y_{ij}}{\sum_{j=1}^{G}\left[\eta_j^{(t)}(1-\tau_{cj}^{(t)}) + (1-\eta_j^{(t)})(1-\nu_{1j}^{(t)})\right]},
\end{aligned}$$

and

$$\begin{aligned}
\sigma_{1i}^{2*(t+1)} &= \frac{\sum_{j=1}^{G}\left[\eta_j^{(t)}\tau_{cj}^{(t)} + (1-\eta_j^{(t)})\nu_{1j}^{(t)}\right]\left(y_{ij}-\mu_{1i}^{*(t+1)}\right)^2}{\sum_{j=1}^{G}\left[\eta_j^{(t)}\tau_{cj}^{(t)} + (1-\eta_j^{(t)})\nu_{1j}^{(t)}\right]}, \\
\sigma_{2i}^{2*(t+1)} &= \frac{\sum_{j=1}^{G}\left[\eta_j^{(t)}(1-\tau_{cj}^{(t)}) + (1-\eta_j^{(t)})(1-\nu_{1j}^{(t)})\right]\left(y_{ij}-\mu_{2i}^{*(t+1)}\right)^2}{\sum_{j=1}^{G}\left[\eta_j^{(t)}(1-\tau_{cj}^{(t)}) + (1-\eta_j^{(t)})(1-\nu_{1j}^{(t)})\right]},
\end{aligned}$$

where we can interpret the denominators of $(\mu_{1i}^*, \sigma_{1i}^{2*})$ and $(\mu_{2i}^*, \sigma_{2i}^{2*})$ as the expected numbers of methylated genes and non-methylated genes respectively. Notice that the above are just weighted averages of the mean and variance formulas of the previous section:

$$\begin{aligned}
\mu_{1i}^{*(t+1)} &= \lambda_1\mu_{1i}^{(t+1)} + (1-\lambda_1)\alpha_{1i}^{(t+1)}, \\
\mu_{2i}^{*(t+1)} &= \lambda_2\mu_{2i}^{(t+1)} + (1-\lambda_2)\alpha_{2i}^{(t+1)}, \\
\sigma_{1i}^{2*(t+1)} &= \lambda_1\sigma_{1i}^{2(t+1)} + (1-\lambda_1)\varsigma_{1i}^{2(t+1)}, \\
\sigma_{2i}^{2*(t+1)} &= \lambda_2\sigma_{2i}^{2(t+1)} + (1-\lambda_2)\varsigma_{2i}^{2(t+1)},
\end{aligned}$$

where we define $\lambda_1 = (\sum_j \eta_j^{(t)}\tau_{cj}^{(t)})/(\sum_j \left[\eta_j^{(t)}\tau_{cj}^{(t)} + (1-\eta_j^{(t)})\nu_{1j}^{(t)}\right])$, and $\lambda_2 = (\sum_j \eta_j^{(t)}(1-\tau_{cj}^{(t)}))/(\sum_j \left[\eta_j^{(t)}(1-\tau_{cj}^{(t)}) + (1-\eta_j^{(t)})(1-\nu_{1j}^{(t)})\right])$.

8.3 Asymptotics

In this section we will explore some large sample properties of the posterior expectations, $E[\gamma_j|\mathbf{y},\boldsymbol{\theta}]$ and $E[w_{cj}|\gamma_j,\mathbf{y},\boldsymbol{\theta}]$ as the number of patients within clusters goes to ∞. By predicting the posterior expectations we hope to retrieve the discriminating genes (genes j with $w_{cj} \neq w_{c'j}$ for some distinct pair $c \neq c'$). Theorem 8.3.1 states that in the limit $E[\gamma_j|\mathbf{y},\boldsymbol{\theta}]$ is equal to 1 when gene j truly discriminates between clusters. We also want to correctly predict the cluster specific methylation patterns for the genes that are declared discriminating. Theorem 8.3.2 implies that conditioning on $\gamma_j = 1$ we correctly predict the methylation status of gene j in each cluster c. Let us now show some desirable properties of $E[\gamma_j|\mathbf{y},\boldsymbol{\theta}]$ in the limit as all cluster sizes tend to infinity. In what follows we assume equal means and variances of the discriminating and non-discriminating density, i.e. $(\mu_{1i},\sigma^2_{1i},\mu_{2i},\sigma^2_{2i}) = (\alpha_{1i},\varsigma^2_{1i},\alpha_{2i},\varsigma^2_{2i})$.

Theorem 8.3.1. *In section 8.2 we derived the posterior expectation of γ_j (see (8.10)). At the true parameters of the model we have*

$$E[\gamma_j|\mathbf{y},\boldsymbol{\theta}] = \frac{pf_1(\mathbf{y}_j)}{pf_1(\mathbf{y}_j)+(1-p)f_2(\mathbf{y}_j)},$$

where

$$f_1(\mathbf{y}_j) = \prod_{c\in\mathcal{C}}\Big(\pi_{1c}\prod_{i\in c}\phi(y_{ij}|\mu_{1i},\sigma^2_{1i}) + \pi_{0c}\prod_{i\in c}\phi(y_{ij}|\mu_{2i},\sigma^2_{2i})\Big),$$

and

$$f_2(\mathbf{y}_j) = \pi_1\prod_{i=1}^{n}\phi(y_{ij}|\mu_{1i},\sigma^2_{1i}) + \pi_0\prod_{i=1}^{n}\phi(y_{ij}|\mu_{2i},\sigma^2_{2i}).$$

Assume the same regularity conditions as in Theorem 7.5.1:

(R.1) The variances are bounded from zero and infinity, i.e. there exist some $\lambda_1,\lambda_2 > 0$ such that $\lambda_1^2 < \sigma^2_{1i},\sigma^2_{2i} < \lambda_2^2$ for all i.

(R.2) The average squared mean differences within each cluster, c, are bounded away from zero and infinity, i.e. $\sup \frac{1}{n_c} \sum_{i=1}^{n_c} \Delta_i^2 < \infty$ *and* $\inf \frac{1}{n_c} \sum_{i=1}^{n_c} \Delta_i^2 > 0$*, where* $\Delta_i = \mu_{2i} - \mu_{1i} > 0$.

(i) Conditioning on $w_{cj} \neq w_{c'j}$ for at least two distinct $c \neq c'$ we have

$$\lim_{n_1,\ldots,n_K \to \infty} E[\gamma_j | \mathbf{y}, \boldsymbol{\theta}] = 1 \quad a.s. \tag{8.19}$$

(ii) Conditioning on $w_{cj} = w_{c'j}$ for all c, c' we have

$$\lim_{n_1,\ldots,n_K \to \infty} E[\gamma_j | \mathbf{y}, \boldsymbol{\theta}] = \begin{cases} \frac{p \prod_{c \in \mathcal{C}} \pi_{1c}}{p \prod_{c \in \mathcal{C}} \pi_{1c} + (1-p)\pi_1} & \text{if } w_{cj} = 1 \text{ for all } c \\ \frac{p \prod_{c \in \mathcal{C}} \pi_{0c}}{p \prod_{c \in \mathcal{C}} \pi_{0c} + (1-p)\pi_0} & \text{if } w_{cj} = 0 \text{ for all } c \end{cases} \quad a.s. \tag{8.20}$$

Note the interpretation of part (ii) above. If we know the true values of w_{cj}, all c, and $w_{cj} = w_{c'j}$ for all c, c' the posterior probability that gene j discriminates between patients is in the limit equal to

$$P(\gamma_j = 1 | w_{cj} = w_{c'j} \text{ all c,c'}) = \begin{cases} \frac{p \prod_{c \in \mathcal{C}} \pi_{1c}}{p \prod_{c \in \mathcal{C}} \pi_{1c} + (1-p)\pi_1} & \text{if } w_{cj} = 1 \text{ for all } c \\ \frac{p \prod_{c \in \mathcal{C}} \pi_{0c}}{p \prod_{c \in \mathcal{C}} \pi_{0c} + (1-p)\pi_0} & \text{if } w_{cj} = 0 \text{ for all } c \end{cases}$$

which is based solely on the prior specifications of $(\gamma_j)_j$ and $(w_{cj})_{c,j}$ given in (8.3) and (8.6).

Proof. For both cases we examine the likelihood in (8.7). Let $\mathcal{C}_1$ denote those clusters c that have $w_{cj} = 1$ and let $\mathcal{C}_2$ denote clusters with $w_{cj} = 0$ and so $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$. If $w_{cj} \neq w_{c'j}$ for at least two distinct $c \neq c'$ then clearly $\mathcal{C}_1 \neq \emptyset$ and $\mathcal{C}_2 \neq \emptyset$. However, we have $\mathcal{C}_1 = \emptyset$ if $w_{cj} = 0$ for all c and $\mathcal{C}_2 = \emptyset$ if $w_{cj} = 1$ for all c. For every cluster $c_k \in \mathcal{C}_k$, $k = 1, 2$, we thus establish that $(y_{ij})_{i \in c_k}$ is a sequence of independent Gaussian variables with means and variances $(\mu_{ki}, \sigma_{ki}^2)_{i \in c_k}$. Note that we can write

$$E[\gamma_j | \mathbf{y}, \boldsymbol{\theta}] = \frac{p}{p + (1-p)(f_2(\mathbf{y}_j)/f_1(\mathbf{y}_j))},$$

and so all we need to show is that

(i) when $w_{cj} \neq w_{c'j}$ for at least two distinct $c \neq c'$

$$\lim_{n_1,\ldots,n_K \to \infty} \frac{f_2(\mathbf{y}_j)}{f_1(\mathbf{y}_j)} = 0 \quad a.s. \tag{8.21}$$

(ii) when $w_{cj} = 1$ for all c

$$\lim_{n_1,\ldots,n_K \to \infty} \frac{f_2(\mathbf{y}_j)}{f_1(\mathbf{y}_j)} = \frac{\pi_1}{\prod_{c \in \mathcal{C}} \pi_{1c}} \quad a.s. \tag{8.22}$$

and when $w_{cj} = 0$ for all c

$$\lim_{n_1,\ldots,n_K \to \infty} \frac{f_2(\mathbf{y}_j)}{f_1(\mathbf{y}_j)} = \frac{\pi_0}{\prod_{c \in \mathcal{C}} \pi_{0c}} \quad a.s. \tag{8.23}$$

We write

$$\begin{aligned}
\frac{f_2(\mathbf{y}_j)}{f_1(\mathbf{y}_j)} &= \frac{\pi_1 \prod_{i=1}^{n} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i}) + \pi_0 \prod_{i=1}^{n} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i})}{\prod_{c \in \mathcal{C}} \left(\pi_{1c} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i}) + \pi_{0c} \prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i}) \right)} \\
&= \frac{\pi_1 \prod_{i=1}^{n} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i}) + \pi_0 \prod_{i=1}^{n} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i})}{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \left\{ \pi_{kc} \prod_{i \in c} \phi(y_{ij}|\mu_{ki}, \sigma^2_{ki}) \left(1 + \frac{\pi_{(3-k),c} \prod_{i \in c} \phi(y_{ij}|\mu_{(3-k),i}, \sigma^2_{(3-k),i})}{\pi_{kc} \prod_{i \in c} \phi(y_{ij}|\mu_{ki}, \sigma^2_{ki})} \right) \right\}} \\
&= \left[\frac{1}{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \pi_{kc}} \right] \cdot \frac{\pi_1 \frac{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i})}{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \prod_{i \in c} \phi(y_{ij}|\mu_{ki}, \sigma^2_{ki})} + \pi_0 \frac{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i})}{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \prod_{i \in c} \phi(y_{ij}|\mu_{ki}, \sigma^2_{ki})}}{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \left(1 + \frac{\pi_{(3-k),c}}{\pi_{kc}} \frac{\prod_{i \in c} \phi(y_{ij}|\mu_{(3-k),i}, \sigma^2_{(3-k),i})}{\prod_{i \in c} \phi(y_{ij}|\mu_{ki}, \sigma^2_{ki})} \right)}.
\end{aligned}$$

For each $k = 1, 2$ and any $c \in \mathcal{C}_k$ we know $(y_{ij})_{i \in c}$ is a sequence of independent Gaussian variables with means and variances $(\mu_{ki}, \sigma^2_{ki})_{i \in c}$ and so

$$\lim_{n_c \to \infty} \frac{\prod_{i \in c} \phi(y_{ij}|\mu_{(3-k),i}, \sigma^2_{(3-k),i})}{\prod_{i \in c} \phi(y_{ij}|\mu_{ki}, \sigma^2_{ki})} = 0 \quad a.s. \tag{8.24}$$

Thus the denominator of the second fraction above tends to 1 as each of the cluster sizes, $n_1, \ldots, n_K$, tend to infinity. The two ratios in the numerator take on different forms depending on which case we are dealing with.

(i) When $w_{cj} \neq w_{c'j}$ for some $c \neq c'$ we have $\mathcal{C}_1 \neq \emptyset$ and $\mathcal{C}_2 \neq \emptyset$ and we get

$$\frac{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)}{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \prod_{i \in c} \phi(y_{ij}|\mu_{ki}, \sigma_{ki}^2)} = \frac{\prod_{c \in \mathcal{C}_2} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)}{\prod_{c \in \mathcal{C}_2} \prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)}.$$

But for each $c \in \mathcal{C}_2$ the ratio $\prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) / \prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)$ tends to 0 as $n_c \to \infty$ (see proof of theorem 7.5.1) and thus the whole expression tends to 0 as $n_c \to \infty$ for all $c \in \mathcal{C}_2$. Similarly we show that the second fraction,

$$\frac{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)}{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \prod_{i \in c} \phi(y_{ij}|\mu_{ki}, \sigma_{ki}^2)} = \frac{\prod_{c \in \mathcal{C}_1} \prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)}{\prod_{c \in \mathcal{C}_1} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)},$$

tends to 0 as $n_c \to \infty$ for all $c \in \mathcal{C}_1$. From (8.24) and the above it is clear that (8.21) holds.

(ii) When $w_{cj} = w_{c'j} = 1$ for all c, c' we have $\mathcal{C}_2 = \emptyset$ and $\mathcal{C}_1 = \mathcal{C}$. This leads to

$$\frac{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)}{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \prod_{i \in c} \phi(y_{ij}|\mu_{ki}, \sigma_{ki}^2)} = \frac{\prod_{c \in \mathcal{C}_1} \prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)}{\prod_{c \in \mathcal{C}_1} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)},$$

which tends to 0 as $n_c \to \infty$ for all $c \in \mathcal{C}_1$,

$$\frac{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)}{\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \prod_{i \in c} \phi(y_{ij}|\mu_{ki}, \sigma_{ki}^2)} = \frac{\prod_{c \in \mathcal{C}_1} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)}{\prod_{c \in \mathcal{C}_1} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)} = 1,$$

and $\prod_{k=1}^{2} \prod_{c \in \mathcal{C}_k} \pi_{kc} = \prod_{c \in \mathcal{C}} \pi_{1c}$. This, along with (8.24), proves (8.22) and a similar argument shows that (8.23) holds when $w_{cj} = w_{c'j} = 0$ for all c, c'.

This completes the proofs of (8.21)-(8.23), which implies the desired results given in (8.19) and (8.20). □

Let us now show that conditioned on γ_j we correctly predict the cluster specific methylation patterns in the limit as cluster sizes tend to infinity.

Theorem 8.3.2. *In section 8.2 we derived the posterior expectation of w_{cj}, given γ_j, (see (8.9)). At the true parameters of the model we have*

$$\begin{aligned} E[w_{cj}|\boldsymbol{\gamma}_j, \mathbf{y}, \boldsymbol{\theta}] &= \gamma_j \frac{\pi_{1c} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i})}{\pi_{1c} \prod_{i \in c} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i}) + \pi_{0c} \prod_{i \in c} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i})} \\ &\quad + (1-\gamma_j) \frac{\pi_1 \prod_{i=1}^{n} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i})}{\pi_1 \prod_{i=1}^{n} \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i}) + \pi_0 \prod_{i=1}^{n} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i})}. \end{aligned}$$

Assuming the same regularity conditions, (R.1) and (R.2), as in theorem 8.3.1 we have

$$\lim_{n_c \to \infty} E[w_{cj}|\boldsymbol{\gamma}_j, \mathbf{y}, \boldsymbol{\theta}] = w_{cj} \quad a.s.$$

Proof. We consider the two cases when we condition on $\gamma_j = 1$ and $\gamma_j = 0$ separately. The proof of each case is identical to the proof of theorem 7.5.1. □

Remark 8.3.1. *Note that theoretically our model allows for cases where $\gamma_j = 1$ and $w_{cj} = w_{c'j}$ for all c, c'. In fact when we explored the likelihood given in (8.7) we saw that it's not identifable in terms of the parameters, $\boldsymbol{\gamma}$ and $\mathbf{w}$. Ultimately what we are interested in is finding the discriminating genes and we do not care so much about the noisy genes. Theorem 8.3.1 states that when gene j truly is discriminating, $w_{cj} \neq w_{c'j}$, for all c, c', then we correctly predict $E[\gamma_j|\mathbf{y}, \boldsymbol{\theta}] = 1$ in the limit. Furthermore, theorem 8.3.2 states that conditioned on $\gamma_j = 1$ we correctly predict the methylation status of gene j in each cluster and thus through those predictions we should be able to detect when $w_{cj} \neq w_{c'j}$, for some c, c'.*

8.4 Two way Classification EM and gene importance prediction

In this section we generalize the two way classification EM algorithm of section 7.7 to account for all genes. As before we assume the true partition is $\mathcal{C}$, and we define the cluster membership indicators

$$X_{ic} = \begin{cases} 1 & \text{if patient } i \text{ is in cluster } c \\ 0 & \text{otherwise} \end{cases}$$

for all $i = 1, \ldots, n$ and $c \in \mathcal{C}$. The data $\mathbf{y}_1, \ldots, \mathbf{y}_n$ is assumed to be a realization of the density

$$f(\mathbf{y}|\mathbf{X}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \prod_{c\in\mathcal{C}} \prod_{i=1}^{n} \Big(f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\gamma}, \boldsymbol{\theta}_i) \Big)^{X_{ic}}, \tag{8.25}$$

where now the density of the data for patient i on the assumption that he/she is in cluster c is given by

$$\begin{aligned} f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\gamma}, \boldsymbol{\theta}_i) &= \prod_{j=1}^{G} \Big\{ \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)^{w_{cj}} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)^{1-w_{cj}} \Big\}^{\gamma_j} \\ &\quad \times \Big\{ \phi(y_{ij}|\alpha_{1i}, \varsigma_{1i}^2)^{w_{1j}} \phi(y_{ij}|\alpha_{2i}, \varsigma_{2i}^2)^{1-w_{1j}} \Big\}^{1-\gamma_j}. \end{aligned} \tag{8.26}$$

Just as in section 7.7 we assume that the methylation indicators, $(w_{cj})_{c,j}$ are fixed, and we assume the same for the gene importance indicators, $(\gamma_j)_j$. However, we assume that the cluster membership indicators, $\mathbf{X}_i$, $i = 1, \ldots, n$, are independent latent random variables that follow a Multinom$(1, \mathbf{p})$ distribution and so

$$f(\mathbf{X}) = \prod_{i=1}^{n} \prod_{c\in\mathcal{C}} p_c^{X_{ic}}, \tag{8.27}$$

where $\mathbf{p} = (p_1, \ldots, p_K)$. We see that

$$f(\mathbf{y}|\mathbf{X}) f(\mathbf{X}) = \prod_{i=1}^{n} \prod_{c\in\mathcal{C}} \Big(p_c f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\gamma}, \boldsymbol{\theta}_i) \Big)^{X_{ic}},$$

and after integrating out $\mathbf{X}$ we obtain the marginal likelihood

$$f(\mathbf{y}) = \prod_{i=1}^{n} \sum_{c\in\mathcal{C}} p_c f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\gamma}, \boldsymbol{\theta}_i).$$

From the above it is clear that the posterior distribution of $\mathbf{X}$ is that of independent multinomials

$$f(\mathbf{X}|\mathbf{y}) = \prod_{i=1}^{n}\prod_{c\in\mathcal{C}}\left(\frac{p_c f(\mathbf{y}_i|\mathbf{w}_c,\boldsymbol{\gamma},\boldsymbol{\theta}_i)}{\sum_k p_k f(\mathbf{y}_i|\mathbf{w}_k,\boldsymbol{\gamma},\boldsymbol{\theta}_i)}\right)^{X_{ic}}.$$

Note that the likelihood in (8.25) can be written as

$$\begin{aligned} f(\mathbf{y}|\mathbf{X}) &= \prod_{c\in\mathcal{C}}\prod_{j=1}^{G}\left(\Big\{\prod_{i=1}^{n}\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2)^{X_{ic}}\Big\}^{w_{cj}}\Big\{\prod_{i=1}^{n}\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2)^{X_{ic}}\Big\}^{1-w_{cj}}\right)^{\gamma_j} \\ &\quad\times\prod_{j=1}^{G}\left(\Big\{\prod_{i=1}^{n}\phi(y_{ij}|\alpha_{1i},\varsigma_{1i}^2)\Big\}^{w_{1j}}\Big\{\prod_{i=1}^{n}\phi(y_{ij}|\alpha_{2i},\varsigma_{2i}^2)\Big\}^{1-w_{1j}}\right)^{1-\gamma_j}, \quad (8.28) \end{aligned}$$

and so the complete data loglikelihood becomes

$$\begin{aligned} \ell_{comp}(\mathbf{y},\mathbf{X}) &= \sum_{c\in\mathcal{C}}\sum_{i=1}^{n}X_{ic}\log p_c \\ &\quad+\sum_{c\in\mathcal{C}}\sum_{j=1}^{G}\gamma_j w_{cj}\sum_{i=1}^{n}X_{ic}\log\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2) \\ &\quad+\sum_{c\in\mathcal{C}}\sum_{j=1}^{G}\gamma_j(1-w_{cj})\sum_{i=1}^{n}X_{ic}\log\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2) \\ &\quad+\sum_{j=1}^{G}(1-\gamma_j)w_{1j}\sum_{i=1}^{n}\log\phi(y_{ij}|\alpha_{1i},\varsigma_{1i}^2) \\ &\quad+\sum_{j=1}^{G}(1-\gamma_j)(1-w_{1j})\sum_{i=1}^{n}\log\phi(y_{ij}|\alpha_{2i},\varsigma_{2i}^2). \end{aligned}$$

8.4.1 E-step

At a current iterate of the parameter estimates $(\gamma_j^{(t)})_j$, $(w_{cj}^{(t)})_{c,j}$, $(\boldsymbol{\theta}_i^{(t)})_i$ and $(p_c^{(t)})_c$ we define the posterior expectation of X_{ic} as

$$\kappa_{ic}^{(t)} = \frac{p_c^{(t)} f(\mathbf{y}_i|\mathbf{w}_c^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{\theta}_i^{(t)})}{\sum_k p_k^{(t)} f(\mathbf{y}_i|\mathbf{w}_k^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{\theta}_i^{(t)})}, \quad (8.29)$$

and so the Q-function has the following form

$$\begin{aligned} Q(\mathbf{\Lambda}|\mathbf{\Lambda}^{(t)}) &= \sum_{c\in\mathcal{C}}\sum_{i=1}^{n}\kappa_{ic}^{(t)}\log p_c \\ &+\sum_{c\in\mathcal{C}}\sum_{j=1}^{G}\gamma_j w_{cj}\sum_{i=1}^{n}\kappa_{ic}^{(t)}\log\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2) \\ &+\sum_{c\in\mathcal{C}}\sum_{j=1}^{G}\gamma_j(1-w_{cj})\sum_{i=1}^{n}\kappa_{ic}^{(t)}\log\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2) \\ &+\sum_{j=1}^{G}(1-\gamma_j)w_{1j}\sum_{i=1}^{n}\log\phi(y_{ij}|\alpha_{1i},\varsigma_{1i}^2) \\ &+\sum_{j=1}^{G}(1-\gamma_j)(1-w_{1j})\sum_{i=1}^{n}\log\phi(y_{ij}|\alpha_{2i},\varsigma_{2i}^2), \end{aligned} \tag{8.30}$$

where $\mathbf{\Lambda} = \{(\gamma_j)_j, (w_{cj})_{c,j}, (\mu_{1i},\sigma_{1i}^2,\mu_{2i},\sigma_{2i}^2,\mu_{1i},\sigma_{1i}^2,\mu_{2i},\sigma_{2i}^2)_i, (p_c)_c\}$. In the M-step we maximize the Q-function above with respect to $\mathbf{\Lambda}$, which requires an iterative (classification EM) algorithm.

8.4.2 Maximizing the Q-function

Note that at a current value of $\kappa_{ic}^{(t)}$ the function $Q(\mathbf{\Lambda}|\mathbf{\Lambda}^{(t)})$ given in (8.30) is of similar form to the Q-function given by (8.11) in section 8.2, where $(\gamma_j)_j$ and $(w_{cj})_{c,j}$ are assumed random bernoullis. The EM algorithm of section 8.2 involved iterating between maximizing the Q-function with respect to the parameters and updating the γ_j and w_{cj} with posterior expectations. However, as explained in section 7.7, when we assume these indicators are fixed rather than random, we maximize (8.30) with a modified (classification) EM algorithm. The cluster proportions maximizer is simply

$$p_c^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\kappa_{ic}^{(t)} \tag{8.31}$$

but in order to maximize with respect to $(\gamma_j)_j$, $(w_{cj})_{c,j}$ and the means and variances we need to iterate between these two steps

1. For given values of $(\gamma_j)_j$ and $(w_{cj})_{c,j}$ we differentiate (8.30) with respect to the individual specific parameters to obtain the maximizers

$$\hat{\mu}_{1i} = \frac{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} \gamma_j w_{cj} y_{ij}}{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} \gamma_j w_{cj}}, \tag{8.32}$$

$$\hat{\mu}_{2i} = \frac{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} \gamma_j (1-w_{cj}) y_{ij}}{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} \gamma_j (1-w_{cj})}, \tag{8.33}$$

$$\hat{\sigma}_{1i}^2 = \frac{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} \gamma_j w_{cj} (y_{ij}-\hat{\mu}_{1i})^2}{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} \gamma_j w_{cj}}, \tag{8.34}$$

$$\hat{\sigma}_{2i}^2 = \frac{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} \gamma_j (1-w_{cj}) (y_{ij}-\hat{\mu}_{2i})^2}{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} \gamma_j (1-w_{cj})}. \tag{8.35}$$

and

$$\hat{\alpha}_{1i} = \frac{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} (1-\gamma_j) w_{1j} y_{ij}}{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} (1-\gamma_j) w_{1j}}, \tag{8.36}$$

$$\hat{\alpha}_{2i} = \frac{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} (1-\gamma_j)(1-w_{1j}) y_{ij}}{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} (1-\gamma_j)(1-w_{1j})}, \tag{8.37}$$

$$\hat{\varsigma}_{1i}^2 = \frac{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} (1-\gamma_j) w_{1j} (y_{ij}-\hat{\alpha}_{1i})^2}{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} (1-\gamma_j) w_{1j}}, \tag{8.38}$$

$$\hat{\varsigma}_{2i}^2 = \frac{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} (1-\gamma_j)(1-w_{1j}) (y_{ij}-\hat{\alpha}_{2i})^2}{\sum_{c=1}^{K} \kappa_{ic}^{(t)} \sum_{j=1}^{G} (1-\gamma_j)(1-w_{1j})}. \tag{8.39}$$

Note that in the special case of $(\mu_{1i}, \sigma^2_{1i}, \mu_{2i}, \sigma^2_{2i}) = (\alpha_{1i}, \varsigma^2_{1i}, \alpha_{2i}, \varsigma^2_{2i})$ *the updating formulas for the joint means and variances become weighted averages of the above formulas*:

$$\begin{aligned}
\hat{\mu}^*_{1i} &= \lambda_1 \hat{\mu}_{1i} + (1-\lambda_1)\hat{\alpha}_{1i}, \\
\hat{\mu}^*_{2i} &= \lambda_2 \hat{\mu}_{2i} + (1-\lambda_2)\hat{\alpha}_{2i}, \\
\hat{\sigma}^{2*}_{1i} &= \lambda_1 \hat{\sigma}^2_{1i} + (1-\lambda_1)\hat{\varsigma}^2_{1i}, \\
\hat{\sigma}^{2*}_{2i} &= \lambda_2 \hat{\sigma}^2_{2i} + (1-\lambda_2)\hat{\varsigma}^2_{2i},
\end{aligned}$$

where we defined $\lambda_1 = (\sum_j \gamma_j w_{cj})/(\sum_j \big[\gamma_j w_{cj} + (1-\gamma_j) w_{1j}\big])$, *and* $\lambda_2 = (\sum_j \gamma_j (1-w_{cj}))/(\sum_j \big[\gamma_j (1-w_{cj}) + (1-\gamma_j)(1-w_{1j})\big])$.

2. For given values of $(\mu_{1i}, \sigma^2_{1i}, \mu_{2i}, \sigma^2_{2i}, \alpha_{1i}, \varsigma^2_{1i}, \alpha_{2i}, \varsigma^2_{2i})_i$ and $(\gamma_j)_j$ we see from (8.30) that the maximizing formula for the cluster specific methylation indicators is

$$\hat{w}_{cj} = \gamma_j \cdot \hat{w}_{cj(\boldsymbol{\gamma})} + (1-\gamma_j) \cdot \hat{w}_{1j(\bar{\boldsymbol{\gamma}})}, \tag{8.40}$$

where

$$\begin{aligned}
\hat{w}_{cj(\boldsymbol{\gamma})} &= \mathbb{1}\Big(\sum_{i=1}^n \kappa^{(t)}_{ic} \log \phi(y_{ij} | \mu_{1i}, \sigma^2_{1i}) > \sum_{i=1}^n \kappa^{(t)}_{ic} \log \phi(y_{ij} | \mu_{2i}, \sigma^2_{2i}) \Big), \\
\hat{w}_{1j(\bar{\boldsymbol{\gamma}})} &= \mathbb{1}\Big(\sum_{i=1}^n \log \phi(y_{ij} | \alpha_{1i}, \varsigma^2_{1i}) > \sum_{i=1}^n \log \phi(y_{ij} | \alpha_{2i}, \varsigma^2_{2i}) \Big).
\end{aligned}$$

Let

$$\begin{aligned}
\hat{\lambda}_j(\mathcal{C}) &= \sum_{c \in \mathcal{C}} \hat{w}_{cj(\boldsymbol{\gamma})} \sum_{i=1}^n \kappa^{(t)}_{ic} \log \phi(y_{ij} | \mu_{1i}, \sigma^2_{1i}) \\
&\quad + \sum_{c \in \mathcal{C}} (1 - \hat{w}_{cj(\boldsymbol{\gamma})}) \sum_{i=1}^n \kappa^{(t)}_{ic} \log \phi(y_{ij} | \mu_{2i}, \sigma^2_{2i}), \\
\hat{\lambda}_j([n]) &= \hat{w}_{1j(\bar{\boldsymbol{\gamma}})} \sum_{i=1}^n \log \phi(y_{ij} | \alpha_{1i}, \varsigma^2_{1i}) \\
&\quad + (1 - \hat{w}_{1j(\bar{\boldsymbol{\gamma}})}) \sum_{i=1}^n \log \phi(y_{ij} | \alpha_{2i}, \varsigma^2_{2i}),
\end{aligned}$$

where $[n] = \{1, \ldots, n\}$. If we plug the values of $(\hat{w}_{cj})_{c,j}$ in (8.40) into (8.30) we see that the maximizing formula for the gene importance indicators is

$$\hat{\gamma}_j \;=\; \mathbb{1}\big(\hat{\lambda}_j(\mathcal{C}) > \hat{\lambda}_j([n])\big), \tag{8.41}$$

We now iterate between steps 1. and 2. until convergence is reached and we arrive at the parameter updates $(\mu_{1i}^{(t+1)}, \sigma_{1i}^{2(t+1)}, \mu_{2i}^{(t+1)}, \sigma_{2i}^{2(t+1)}, \alpha_{1i}^{(t+1)}, \varsigma_{1i}^{2(t+1)}, \alpha_{2i}^{(t+1)}, \varsigma_{2i}^{2(t+1)})_i$, $(\gamma_j^{(t+1)})_j$ and $(w_{cj}^{(t+1)})_{c,j}$.

Remark 8.4.1. *Note that when $(\mu_{1i}, \sigma_{1i}^2, \mu_{2i}, \sigma_{2i}^2) = (\alpha_{1i}, \varsigma_{1i}^2, \alpha_{2i}, \varsigma_{2i}^2)$ the resulting MLEs of the latent indicators have some nice properties. We thus recommend this assumption, both for biological reasons and for the reasons we'll list below. When the means and variances are not forced to be equal we can get estimates for a particular gene, j, that has $\hat{\gamma}_j = 1$ but $\hat{w}_{cj} = \hat{w}_{c'j}$ for all c, c'. This is not necessarily unreasonable as we have discussed in Remark 8.1.1. However, in the two way classification algorithm we are guaranteed not to arrive at estimates such as just described if we assume $(\mu_{1i}, \sigma_{1i}^2, \mu_{2i}, \sigma_{2i}^2) = (\alpha_{1i}, \varsigma_{1i}^2, \alpha_{2i}, \varsigma_{2i}^2)$. More specifically, if gene j is not discriminating between the classes, i.e. $\hat{w}_{cj(\boldsymbol{\gamma})} = \hat{w}_{c'j(\boldsymbol{\gamma})}$ for all c, c', the importance indicator, $\hat{\gamma}_j$, in (8.41) is equal to 0. Assume $(\mu_{1i}, \sigma_{1i}^2, \mu_{2i}, \sigma_{2i}^2) = (\alpha_{1i}, \varsigma_{1i}^2, \alpha_{2i}, \varsigma_{2i}^2)$ and $\hat{w}_{cj(\boldsymbol{\gamma})} = \hat{w}_{c'j(\boldsymbol{\gamma})} = 1$ for all c, c'. That implies for all c:*

$$\sum_{i=1}^{n} \kappa_{ic}^{(t)} \log \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) > \sum_{i=1}^{n} \kappa_{ic}^{(t)} \log \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2),$$

which implies

$$\begin{aligned} \sum_{i=1}^{n} \log \phi(y_{ij} | \alpha_{1i}, \varsigma_{1i}^2) \;&=\; \sum_{c} \sum_{i=1}^{n} \kappa_{ic}^{(t)} \log \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) \\ &>\; \sum_{c} \sum_{i=1}^{n} \kappa_{ic}^{(t)} \log \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2) \\ &=\; \sum_{i=1}^{n} \log \phi(y_{ij} | \alpha_{2i}, \varsigma_{2i}^2). \end{aligned} \tag{8.42}$$

But this implies that $\hat{w}_{1j(\bar{\gamma})} = 1$*, and so* $\hat{w}_{cj} = 1$ *from (8.40). This leads to*

$$\begin{aligned}\hat{\lambda}_j(\mathcal{C}) &= \sum_{c\in\mathcal{C}}\sum_{i=1}^{n} \kappa_{ic}^{(t)} \log \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \\ &= \sum_{i=1}^{n} \log \phi(y_{ij}|\alpha_{1i}, \varsigma_{1i}^2) \\ &= \hat{\lambda}_j([n]),\end{aligned}$$

which implies $\hat{\gamma}_j = 0$*. Similar argument holds with inequalities reversed when we assume* $\hat{w}_{cj(\boldsymbol{\gamma})} = \hat{w}_{c'j(\boldsymbol{\gamma})} = 0$ *for all* c, c'*.*

8.5 Classification

The construction of a likelihood for any given partition of the patient set also provides a powerful tool for classification. Assume we have data on n patients and we know which class each patient belongs to. We run the EM algorithm of section 8.2 on the data, $\mathbf{Y}$, to maximize the observed data likelihood given in (8.5). In the process we obtain posterior expectations of the latent indicators, round them to either 0 or 1, and call them $(\hat{\gamma}_j)_j$, and $(\hat{w}_{cj})_{c,j}$. Classification of a new patient i involves treating these rounded posterior expectations as the true values. By looking at (8.7) we see that given the gene importance and methylation indicators the conditional likelihood of a new observation $(y_{ij})_j$, on the assumption that $i \in c$, is given by

$$\begin{aligned}L_c &= \prod_{j=1}^{G} \left\{ \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)^{\hat{w}_{cj}} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)^{1-\hat{w}_{cj}} \right\}^{\hat{\gamma}_j} \\ &\quad \times \left\{ \phi(y_{ij}|\alpha_{1i}, \varsigma_{1i}^2)^{\hat{w}_{1j}} \phi(y_{ij}|\alpha_{2i}, \varsigma_{2i}^2)^{1-\hat{w}_{1j}} \right\}^{1-\hat{\gamma}_j}.\end{aligned} \tag{8.43}$$

We want to assign patient $i \in c$ if $L_c > L_{c'}$ for all $c' \neq c$. But note that the latter part does not involve c and so if we define $\hat{J}_d(\mathcal{C}) = \{j|\hat{\gamma}_j = 1\}$ we arrive at the discriminant

likelihood

$$L_c(\mathbf{y}_i|\mathbf{Y},\boldsymbol{\theta}_i) = \prod_{j\in\hat{J}_d(\mathcal{C})} \phi(y_{ij}|\mu_{1i},\sigma^2_{1i})^{\hat{w}_{cj}}\phi(y_{ij}|\mu_{2i},\sigma^2_{2i})^{1-\hat{w}_{cj}}. \tag{8.44}$$

The above is maximized at

$$\begin{aligned}\hat{\mu}_{1i} &= \frac{\sum_{j\in\hat{J}_d(\mathcal{C})}\hat{w}_{cj}y_{ij}}{\sum_{j\in\hat{J}_d(\mathcal{C})}\hat{w}_{cj}},\\ \hat{\sigma}^2_{1i} &= \frac{\sum_{j\in\hat{J}_d(\mathcal{C})}\hat{w}_{cj}\big(y_{ij}-\hat{\mu}_{1i}\big)^2}{\sum_{j\in\hat{J}_d(\mathcal{C})}\hat{w}_{cj}},\end{aligned}$$

and

$$\begin{aligned}\hat{\mu}_{2i} &= \frac{\sum_{j\in\hat{J}_d(\mathcal{C})}(1-\hat{w}_{cj})y_{ij}}{\sum_{j\in\hat{J}_d(\mathcal{C})}(1-\hat{w}_{cj})},\\ \hat{\sigma}^2_{2i} &= \frac{\sum_{j\in\hat{J}_d(\mathcal{C})}(1-\hat{w}_{cj})\big(y_{ij}-\hat{\mu}_{2i}\big)^2}{\sum_{j\in\hat{J}_d(\mathcal{C})}(1-\hat{w}_{cj})}.\end{aligned}$$

Note that the above estimates are explicit functions of the estimated cluster specific indicators. By plugging these estimates into (8.44) we arrive at the following discriminant rule

$$i\in c \quad \text{if} \quad L_c(\mathbf{y}_i|\mathbf{Y},\hat{\boldsymbol{\theta}}_i(\hat{\mathbf{w}}_c)) > L_{c'}(\mathbf{y}_i|\mathbf{Y},\hat{\boldsymbol{\theta}}_i(\hat{\mathbf{w}}_{c'})) \text{ for all } c'\neq c.$$

Empirical discriminant analysis suggests that a better classification success rate is achieved by plugging the above mean and variance estimates into the following refined likelihood

$$L^A_c = \prod_{j\in\hat{J}_d(\mathcal{C})\cap\hat{J}_A} \phi(y_{ij}|\mu_{1i},\sigma^2_{1i})^{\hat{w}_{cj}}\phi(y_{ij}|\mu_{2i},\sigma^2_{2i})^{1-\hat{w}_{cj}}.$$

where $\hat{J}_A = \{j|\hat{w}_{cj}\neq\hat{w}_{c'j}$ for at least one pair $c\neq c'\}$ and then let $i\in c$ if $L^A_c > L^A_{c'}$ all $c'\neq c$.

8.5.1 Approximate Bayesian approach

In this section we will argue that the above method can be thought of as an approximate Bayesian discriminant rule. In the Bayesian framework we look at the density of $\mathbf{y}_i$, on

the assumption that $i \in c$,

$$f_c(\mathbf{y}_i|\mathbf{w},\boldsymbol{\gamma},\boldsymbol{\theta}_i) = \prod_{j=1}^{G}\left\{\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2)^{w_{cj}}\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2)^{1-w_{cj}}\right\}^{\gamma_j} \times\left\{\phi(y_{ij}|\alpha_{1i},\varsigma_{1i}^2)^{w_{1j}}\phi(y_{ij}|\alpha_{2i},\varsigma_{2i}^2)^{1-w_{1j}}\right\}^{1-\gamma_j}, \quad (8.45)$$

and have priors $\pi(\mathbf{w}|\boldsymbol{\gamma})$ and $\pi(\boldsymbol{\gamma})$. Notice the difference between this setup and the usual Bayesian classification setup. Here we have additional individual specific parameters, $\boldsymbol{\theta}_i$, for the new patient which we can think of as adjustment parameters to account for the micro array effect. Given the data, $\mathbf{Y}$, the likelihood of a new observation, $\mathbf{y}_i$, on the assumption that $i \in c$, is obtained by averaging the p.d.f. in (8.45) with respect to the posterior distribution $f(\mathbf{w},\boldsymbol{\gamma}|\mathbf{Y}) \propto f(\mathbf{Y}|\mathbf{w},\boldsymbol{\gamma})\pi(\mathbf{w}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})$. This results in

$$L_c(\mathbf{y}_i|\mathbf{Y},\boldsymbol{\theta}_i) = \sum_{\mathbf{w}}\sum_{\boldsymbol{\gamma}} f_c(\mathbf{y}_i|\mathbf{w},\boldsymbol{\gamma},\boldsymbol{\theta}_i)f(\mathbf{w}|\boldsymbol{\gamma},\mathbf{Y})f(\boldsymbol{\gamma}|\mathbf{Y}). \quad (8.46)$$

In the usual Bayesian setup the above expression would not depend on any unknown parameters and we would allocate patient i to the cluster corresponding to the largest value of (8.46). To account for the individual specific microarray adjustment parameter, $\boldsymbol{\theta}_i$, we can simply maximize L_c with respect to $\boldsymbol{\theta}_i$ and let $i \in c$ if $\hat{L}_c > \hat{L}_{c'}$ for all $c' \neq c$. It is easy to see that it is intractable to maximize the expression given in (8.46) with respect to the individual specific means and variances. However, an approximation to the above average is obtained by simply plugging the posterior expectations $E[w_{cj}|\gamma_j,\mathbf{Y}]$ and $E[\gamma_j|\mathbf{Y}]$ into $f_c(\mathbf{y}_i|\mathbf{w},\boldsymbol{\gamma},\boldsymbol{\theta}_i)$ to obtain

$$L_c = \prod_{j=1}^{G}\left\{\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2)^{\hat{w}_{cj}}\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2)^{1-\hat{w}_{cj}}\right\}^{\hat{\gamma}_j} \times\left\{\phi(y_{ij}|\alpha_{1i},\varsigma_{1i}^2)^{\hat{w}_{1j}}\phi(y_{ij}|\alpha_{2i},\varsigma_{2i}^2)^{1-\hat{w}_{1j}}\right\}^{1-\hat{\gamma}_j},$$

but this is simply the expression given in (8.43).

8.5.2 Discriminant rule, equal means and variances

Now let us explore how the discriminant rule looks like in the special case when $(\mu_{1i}, \sigma^2_{1i}, \mu_{2i}, \sigma^2_{2i}) = (\alpha_{1i}, \varsigma^2_{1i}, \alpha_{2i}, \varsigma^2_{2i})$ for all i. We first run the EM algorithm of section 8.2 on the data, $\mathbf{Y}$, using the M-step of subsection 8.2.3. As before we obtain posterior expectations of the latent indicators, $(\gamma_j)_j$ and $(w_{cj})_{c,j}$ and treat them as if they were the true values of the indicators. The conditional likelihood of a new observation $(y_{ij})_j$, on the assumption $i \in c$, is given by

$$\begin{aligned} L_c &= \prod_{j=1}^{G} \Big\{ \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i})^{w_{cj}} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i})^{1-w_{cj}} \Big\}^{\gamma_j} \\ &\quad \times \Big\{ \phi(y_{ij}|\mu_{1i}, \sigma^2_{1i})^{w_{1j}} \phi(y_{ij}|\mu_{2i}, \sigma^2_{2i})^{1-w_{1j}} \Big\}^{1-\gamma_j}, \end{aligned} \tag{8.47}$$

but now the means and variances all depend on c and so we cannot dismiss the latter part as before. The above is maximized at

$$\begin{aligned} \mu^*_{1i} &= \frac{\sum_{j=1}^{G} \Big[\gamma_j w_{cj} + (1-\gamma_j) w_{1j} \Big] y_{ij}}{\sum_{j=1}^{G} \Big[\gamma_j w_{cj} + (1-\gamma_j) w_{1j} \Big]}, \\ \mu^*_{2i} &= \frac{\sum_{j=1}^{G} \Big[\gamma_j (1-w_{cj}) + (1-\gamma_j)(1-w_{1j}) \Big] y_{ij}}{\sum_{j=1}^{G} \Big[\gamma_j (1-w_{cj}) + (1-\gamma_j)(1-w_{1j}) \Big]}, \end{aligned}$$

and

$$\begin{aligned} \sigma^{2*}_{1i} &= \frac{\sum_{j=1}^{G} \Big[\gamma_j w_{cj} + (1-\gamma_j) w_{1j} \Big] \big(y_{ij} - \mu^*_{1i} \big)^2}{\sum_{j=1}^{G} \Big[\gamma_j w_{cj} + (1-\gamma_j) w_{1j} \Big]}, \\ \sigma^{2*}_{2i} &= \frac{\sum_{j=1}^{G} \Big[\gamma_j (1-w_{cj}) + (1-\gamma_j)(1-w_{1j}) \Big] \big(y_{ij} - \mu^*_{2i} \big)^2}{\sum_{j=1}^{G} \Big[\gamma_j (1-w_{cj}) + (1-\gamma_j)(1-w_{1j}) \Big]}, \end{aligned}$$

and upon plugging these into (8.47) we let $i \in c$ if $L_c > L_{c'}$ for all $c' \neq c$. Empirical discriminant analysis suggests that a better classification success rate is achieved by plugging the above means and variance estimates into the following refined likelihood

$$L_c^A = \prod_{j \in \hat{J}_d(\mathcal{C}) \cap \hat{J}_A} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)^{\hat{w}_{cj}} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)^{1-\hat{w}_{cj}},$$

where $\hat{J}_d(\mathcal{C}) = \{j|\gamma_j = 1\}$ and $\hat{J}_A = \{j|\hat{w}_{cj} \neq \hat{w}_{c'j}$ for at least one pair $c \neq c'\}$. This makes sense as here we are only comparing the likelihoods on genes that actually methylate differently across classes.

8.6 Multiple platforms

As discussed in section 7.6 our likelihood based model can be extended to account for multiple data types as long as each data type can be reasonably modeled with the above methods. We now extend the model above in a straightforward manner to account for multiple data platforms.

8.6.1 The extended partition likelihood on multiple platforms

In what follows we use the notation of section 7.6 and recall the definition of the methylation indicators

$$w_{cjk} = \begin{cases} 1 & \text{if DNA fragment } j \text{ on platform } k \text{ is ON in cluster } c \\ 0 & \text{if DNA fragment } j \text{ on platform } k \text{ is OFF in cluster } c \end{cases}$$

We also introduce a platform specific DNA fragment importance indicator,

$$\gamma_{jk} = \begin{cases} 1 & \text{if DNA fragment } j \text{ on platform } k \text{ discriminates between patients} \\ 0 & \text{otherwise} \end{cases}$$

The likelihood of the data is identical to the likelihood given in (8.7) except we now have a product over the platforms $k = 1, \ldots, m$

$$
\begin{aligned}
& f(\mathbf{y}|\mathbf{w}, \boldsymbol{\gamma}) \\
= & \prod_{c \in \mathcal{C}} \prod_{i \in c} f(\mathbf{y}_i | \mathbf{w}_c, \boldsymbol{\gamma}, \boldsymbol{\theta}_i), \\
= & \prod_{c \in \mathcal{C}} \prod_{i \in c} \prod_{k=1}^{m} \prod_{j=1}^{G_k} \Big\{ \phi(y_{ijk} | \mu_{1ik}, \sigma^2_{1ik})^{w_{cjk}} \phi(y_{ijk} | \mu_{2ik}, \sigma^2_{2ik})^{1-w_{cjk}} \Big\}^{\gamma_{jk}} \\
& \qquad \times \Big\{ \phi(y_{ijk} | \alpha_{1ik}, \varsigma^2_{1ik})^{w_{1jk}} \phi(y_{ijk} | \alpha_{2ik}, \varsigma^2_{2ik})^{1-w_{1jk}} \Big\}^{1-\gamma_{jk}}. \\
= & \prod_{k=1}^{m} \prod_{j=1}^{G_k} \left(\prod_{c \in \mathcal{C}} \Big\{ \prod_{i \in c} \phi(y_{ijk} | \mu_{1ik}, \sigma^2_{1ik}) \Big\}^{w_{cjk}} \Big\{ \prod_{i \in c} \phi(y_{ijk} | \mu_{2ik}, \sigma^2_{2ik}) \Big\}^{1-w_{cjk}} \right)^{\gamma_{jk}} \\
& \times \left(\Big\{ \prod_{i=1}^{n} \phi(y_{ijk} | \alpha_{1ik}, \varsigma^2_{1ik}) \Big\}^{w_{1jk}} \Big\{ \prod_{i=1}^{n} \phi(y_{ijk} | \alpha_{2ik}, \varsigma^2_{2ik}) \Big\}^{1-w_{1jk}} \right)^{1-\gamma_{jk}}.
\end{aligned}
$$

We put the following prior on $\boldsymbol{\gamma} = (\gamma_{jk})_{j,k}$

$$
f(\boldsymbol{\gamma}) = \prod_{k=1}^{m} \prod_{j=1}^{G_k} p_k^{\gamma_{jk}} (1 - p_k)^{1-\gamma_{jk}},
$$

where p_k is the proportion of discriminating DNA fragments in platform k. We put the following conditional prior on $\mathbf{w} = (w_{cjk})_{c,j,k}$

$$
f(\mathbf{w}|\boldsymbol{\gamma}) = \prod_{k=1}^{m} \prod_{j=1}^{G_k} \Big(\prod_{c \in \mathcal{C}} \pi_{1ck}^{w_{cjk}} \pi_{0ck}^{1-w_{cjk}} \Big)^{\gamma_{jk}} \Big(I(\mathbf{w}_{jk} \in A_k) \pi_{1k}^{w_{1jk}} \pi_{0k}^{1-w_{1jk}} \Big)^{1-\gamma_{jk}},
$$

where $A_k = \{\mathbf{w}_{jk} | w_{cjk} = w_{c'jk}, \text{ all } c, c' \in \mathcal{C}\}$. We interpret π_{1ck} as the proportion of methylated genes among discriminating genes in platform k and π_{1k} as the proportion of methylated genes among nondiscriminating genes in platform k. It is not hard to see that the multiple platform EM algorithm simply involves fitting the single platform EM algorithm for each platform separately.

8.6.2 Two way classification EM on Multiple platforms

Extending the two way classification EM algorithm to multiple platforms is straightforward. We follow the above section 8.4 closely and simply replace the density given in (8.26) with the following density

$$\begin{aligned} f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\gamma}, \boldsymbol{\theta}_i) &= \prod_{k=1}^{m}\prod_{j=1}^{G_k}\Big\{\phi(y_{ijk}|\mu_{1ik}, \sigma^2_{1ik})^{w_{cjk}}\phi(y_{ijk}|\mu_{2ik}, \sigma^2_{2ik})^{1-w_{cjk}}\Big\}^{\gamma_{jk}} \\ &\quad \times\Big\{\phi(y_{ijk}|\alpha_{1ik}, \varsigma^2_{1ik})^{w_{1jk}}\phi(y_{ijk}|\alpha_{2ik}, \varsigma^2_{2ik})^{1-w_{1jk}}\Big\}^{1-\gamma_{jk}}. \end{aligned}$$

We then proceed exactly in the same manner as in section 8.4 and arrive at the E-step which involves a slightly modified version of the Q-function in (8.30):

$$\begin{aligned} Q(\boldsymbol{\Lambda}|\boldsymbol{\Lambda}^{(n)}) &= \sum_{c\in\mathcal{C}}\sum_{i=1}^{n}\kappa_{ic}^{(n)}\log p_c \\ &\quad + \sum_{k=1}^{m}\sum_{c\in\mathcal{C}}\sum_{j=1}^{G_k}\gamma_{jk}w_{cjk}\sum_{i=1}^{n}\kappa_{ic}^{(n)}\log\phi(y_{ijk}|\mu_{1ik}, \sigma^2_{1ik}) \\ &\quad + \sum_{k=1}^{m}\sum_{c\in\mathcal{C}}\sum_{j=1}^{G_k}\gamma_{jk}(1-w_{cjk})\sum_{i=1}^{n}\kappa_{ic}^{(n)}\log\phi(y_{ijk}|\mu_{2ik}, \sigma^2_{2ik}) \\ &\quad + \sum_{k=1}^{m}\sum_{j=1}^{G_k}(1-\gamma_{jk})w_{1jk}\sum_{i=1}^{n}\log\phi(y_{ijk}|\alpha_{1ik}, \varsigma^2_{1ik}) \\ &\quad + \sum_{k=1}^{m}\sum_{j=1}^{G_k}(1-\gamma_{jk})(1-w_{1jk})\sum_{i=1}^{n}\log\phi(y_{ijk}|\alpha_{2ik}, \varsigma^2_{2ik}), \end{aligned}$$

where

$$\kappa_{ic}^{(n)} = \frac{p_c^{(n)}f(\mathbf{y}_i|\mathbf{w}_c^{(n)}, \boldsymbol{\gamma}^{(n)}, \boldsymbol{\theta}_i^{(n)})}{\sum_k p_k^{(n)}f(\mathbf{y}_i|\mathbf{w}_k^{(n)}, \boldsymbol{\gamma}^{(n)}, \boldsymbol{\theta}_i^{(n)})},$$

The M-step is even more straightforward and simply involves running the M-step procedure of subsection 8.4.2 for each platform separately to obtain the platform specific parameter updates.

8.6.3 Classification on multiple platforms

We run the single platform EM algorithm on the observed data, $\mathbf{Y}_k$, for each platform k separately. In the process we obtain posterior expectations of the latent indicators, round them to either 0 or 1, and call them $(\hat{\gamma}_{jk})_{j,k}$, and $(\hat{w}_{cjk})_{c,j,k}$. The discriminant likelihood of a new observation, $\mathbf{y}_i$, generalizes in a straightforward manner to

$$\begin{aligned} L_c &= \prod_{k=1}^{m}\prod_{j=1}^{G_k}\left\{\phi(y_{ijk}|\mu_{1ik},\sigma^2_{1ik})^{\hat{w}_{cjk}}\phi(y_{ijk}|\mu_{2ik},\sigma^2_{2ik})^{1-\hat{w}_{cjk}}\right\}^{\hat{\gamma}_{jk}} \\ &\quad\times\left\{\phi(y_{ijk}|\mu_{1ik},\sigma^2_{1ik})^{\hat{w}_{1jk}}\phi(y_{ijk}|\mu_{2ik},\sigma^2_{2ik})^{1-\hat{w}_{1jk}}\right\}^{1-\hat{\gamma}_{jk}}. \end{aligned}$$

and after maximizing L_c with respect to $\boldsymbol{\theta}_i$, for all c, we assign $i \in c$ if $\hat{L}_c > \hat{L}_{c'}$ all $c' \neq c$.

CHAPTER 9

RANDOM EFFECTS MODEL

One of the novelties of our clustering algorithm is the inclusion of individual specific parameters, $(\mu_{1i}, \sigma^2_{1i}, \mu^2_{2i}, \sigma^2_{2i})_i$, in the model. This means that the number of parameters grows with the number of patients in the study. This is somewhat unsatisfactory, but since the number of genes is much larger than the number of patients in this setting estimating these parameters is not a problem. However, we could imagine a situation where the number of patients is comparable, or much larger than the number of measurements per patient and in those cases it might be unfeasible to include so many parameters in the model. Since the individual parameters are supposed to account for between subject variability (microarray effect) it seems more natural to treat the individual means and variances as random samples from some distribution. This is a well accepted practice in the mixed model literature and the number of parameters no longer grows with increasing amount of data. In this chapter we present a more realistic model. For simplicity we will explain these ideas on the model given in section 7.1 rather than the extended model of chapter 8. In section 9.1 we specify the model and in section 9.2 we show that in the case when the number of genes is far greater than the number of patients the random effects model can be estimated approximately using the previous methods of chapter 7.

9.1 Random effects model

Given the unknown partition, $\mathcal{C}$, the data $\mathbf{y}_1, \ldots, \mathbf{y}_n$ is assumed to be a realization of the density

$$f(\mathbf{y}|\mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\theta}) = \prod_{c \in \mathcal{C}} \prod_{i \in c} f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\theta}_i), \tag{9.1}$$

where

$$f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\theta}_i) = \prod_{j \in J_d} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)^{w_{cj}} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)^{1-w_{cj}}. \tag{9.2}$$

For ease of notation assume $J_d = \{1, \ldots, G\}$ and that patients in cluster c are labeled $i = 1, \ldots, n_c$. Let $\boldsymbol{\theta}_c = (\boldsymbol{\theta}_i)_{i=1}^{n_c}$ and define the cluster specific likelihood,

$$L(\mathbf{w}_c, \boldsymbol{\theta}_c) = \prod_{i=1}^{n_c} f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\theta}_i). \tag{9.3}$$

Note that the likelihood in (9.1) separates, $L(\mathbf{w}, \boldsymbol{\theta}) = \prod_{c \in \mathcal{C}} L(\mathbf{w}_c, \boldsymbol{\theta}_c)$ and thus for simplicity we only focus on a single cluster c in the following derivation. We treat the means and variances, of patients within cluster c, as random effects and specify the conjugate priors,

$$\begin{aligned} \mu_{ki} &\sim \mathrm{N}(\mu_{kc}, r_{kc}\sigma_{ki}^2), \\ \sigma_{ki}^2 &\sim \mathrm{IG}(\alpha_{kc}, \beta_{kc}), \end{aligned}$$

independently for all $k = 1, 2$ and $i = 1, \ldots, n_c$. Denote the hyper-parameters, $(\mu_{kc}, r_{kc}, \sigma_{ki}^2, \alpha_{kc}, \beta_{kc})_{k=1}^2$, by $\boldsymbol{\psi}_c$. Integrating out the random effects we arrive at the marginal likelihood

$$L_c(\mathbf{w}_c, \boldsymbol{\psi}_c) = \int L(\mathbf{w}_c, \boldsymbol{\theta}_c) \prod_{k=1}^{2} \prod_{i=1}^{n_c} f(\mu_{ki}) f(\sigma_{ki}^2) d(\mu_{ki}) d(\sigma_{ki}^2).$$

For now let us not worry about the parameter $\mathbf{w}_c$ and for sake of clarity assume it is fixed in what follows. From (9.2) and (9.3) we see that

$$\begin{aligned} L(\mathbf{w}_c, \boldsymbol{\theta}_c) &= (2\pi)^{-(Gn_c)/2} \prod_{i=1}^{n_c} (\sigma_{1i}^2)^{-\sum w_{cj}/2} \exp\Big\{ -\frac{1}{2\sigma_{1i}^2} \sum_{j=1}^{G} w_{cj}(y_{ij} - \mu_{1i})^2 \Big\} \\ &\times (2\pi)^{-(Gn_c)/2} \prod_{i=1}^{n_c} (\sigma_{2i}^2)^{-\sum (1-w_{cj})/2} \exp\Big\{ -\frac{1}{2\sigma_{2i}^2} \sum_{j=1}^{G} (1 - w_{cj})(y_{ij} - \mu_{2i})^2 \Big\}, \end{aligned}$$

and we have for all $i = 1, \ldots, n_c$

$$\begin{aligned} f(\mu_{1i}|\sigma_{1i}^2) &= (2\pi)^{-1/2} (r_{1c}\sigma_{1i}^2)^{-1/2} \exp\Big\{ -\frac{1}{2r_{1c}\sigma_{1i}^2} (\mu_{1i} - \mu_{1c})^2 \Big\}, \\ f(\mu_{2i}|\sigma_{2i}^2) &= (2\pi)^{-1/2} (r_{2c}\sigma_{2i}^2)^{-1/2} \exp\Big\{ -\frac{1}{2r_{2c}\sigma_{2i}^2} (\mu_{2i} - \mu_{2c})^2 \Big\}, \end{aligned}$$

and

$$
\begin{aligned}
f(\sigma_{1i}^2) &= \frac{\beta_{1c}^{\alpha_{1c}}}{\Gamma(\alpha_{1c})}(\sigma_{1i}^2)^{-(1+\alpha_{1c})}\exp\Big(-\frac{\beta_{1c}}{\sigma_{1i}^2}\Big),\\
f(\sigma_{2i}^2) &= \frac{\beta_{2c}^{\alpha_{2c}}}{\Gamma(\alpha_{2c})}(\sigma_{2i}^2)^{-(1+\alpha_{2c})}\exp\Big(-\frac{\beta_{2c}}{\sigma_{2i}^2}\Big).
\end{aligned}
$$

We get

$$
\begin{aligned}
& L(\mathbf{w}_c,\boldsymbol{\theta}_c)\prod_{i=1}^{n_c} f(\mu_{1i}|\sigma_{1i}^2)f(\sigma_{1i}^2)f(\mu_{2i}|\sigma_{2i}^2)f(\sigma_{2i}^2)\\
= \;& (2\pi)^{-(G+1)n_c}(r_{1c}r_{2c})^{-n_c/2}\beta_{1c}^{n_c\alpha_{1c}}\Gamma(\alpha_{1c})^{-n_c}\beta_{2c}^{n_c\alpha_{2c}}\Gamma(\alpha_{2c})^{-n_c} \qquad (9.4)\\
& \times\prod_{i=1}^{n_c}(\sigma_{1i}^2)^{-1-\alpha_{1c}-(\sum w_{cj}+1)/2}\exp\Big(-\frac{\beta_{1c}}{\sigma_{1i}^2}\Big)\\
& \times\prod_{i=1}^{n_c}(\sigma_{2i}^2)^{-1-\alpha_{2c}-(\sum(1-w_{cj})+1)/2}\exp\Big(-\frac{\beta_{2c}}{\sigma_{2i}^2}\Big)\\
& \times\prod_{i=1}^{n_c}\exp\left\{-\frac{1}{2\sigma_{1i}^2}\Big(\sum_{j=1}^{G}w_{cj}(y_{ij}-\mu_{1i})^2+\frac{1}{r_{1c}}(\mu_{1i}-\mu_{1c})^2\Big)\right\}\\
& \times\prod_{i=1}^{n_c}\exp\left\{-\frac{1}{2\sigma_{2i}^2}\Big(\sum_{j=1}^{G}(1-w_{cj})(y_{ij}-\mu_{2i})^2+\frac{1}{r_{2c}}(\mu_{2i}-\mu_{2c})^2\Big)\right\}\\
= \;& (2\pi)^{-(G+1)n_c}(r_{1c}r_{2c})^{-n_c/2}\beta_{1c}^{n_c\alpha_{1c}}\Gamma(\alpha_{1c})^{-n_c}\beta_{2c}^{n_c\alpha_{2c}}\Gamma(\alpha_{2c})^{-n_c}\\
& \times\prod_{i=1}^{n_c}(\sigma_{1i}^2)^{-1-\alpha_{1c}-(\sum w_{cj}+1)/2}\exp\Big(-\frac{\beta_{1c}}{\sigma_{1i}^2}-\frac{1}{2\sigma_{1i}^2}(\sum_{j=1}^{G}w_{cj}y_{ij}^2+\frac{\mu_{1c}^2}{r_{1c}})\Big)\\
& \times\prod_{i=1}^{n_c}(\sigma_{2i}^2)^{-1-\alpha_{2c}-(\sum(1-w_{cj})+1)/2}\exp\Big(-\frac{\beta_{2c}}{\sigma_{2i}^2}-\frac{1}{2\sigma_{2i}^2}(\sum_{j=1}^{G}(1-w_{cj})y_{ij}^2+\frac{\mu_{2c}^2}{r_{2c}})\Big)\\
& \times\prod_{i=1}^{n_c}\exp\left\{-\frac{1}{2\sigma_{1i}^2}\Big[(\sum_{j=1}^{G}w_{cj}+\frac{1}{r_{1c}})\mu_{1i}^2-2(\sum_{j=1}^{G}w_{cj}y_{ij}+\frac{\mu_{1c}}{r_{1c}})\mu_{1i}\Big]\right\}\\
& \times\prod_{i=1}^{n_c}\exp\left\{-\frac{1}{2\sigma_{2i}^2}\Big[(\sum_{j=1}^{G}(1-w_{cj})+\frac{1}{r_{2c}})\mu_{2i}^2-2(\sum_{j=1}^{G}(1-w_{cj})y_{ij}+\frac{\mu_{2c}}{r_{2c}})\mu_{2i}\Big]\right\}
\end{aligned}
$$

$$
\begin{aligned}
&= (2\pi)^{-(G+1)n_c}(r_{1c}r_{2c})^{-n_c/2}\beta_{1c}^{n_c\alpha_{1c}}\Gamma(\alpha_{1c})^{-n_c}\beta_{2c}^{n_c\alpha_{2c}}\Gamma(\alpha_{2c})^{-n_c} \\
&\quad\times\prod_{i=1}^{n_c}(\sigma_{1i}^2)^{-1-\alpha_{1c}-(\sum w_{cj}+1)/2}\exp\Big(-\frac{\beta_{1c}}{\sigma_{1i}^2}-\frac{1}{2\sigma_{1i}^2}(\sum_{j=1}^{G}w_{cj}y_{ij}^2+\frac{\mu_{1c}^2}{r_{1c}})\Big) \\
&\quad\times\prod_{i=1}^{n_c}(\sigma_{2i}^2)^{-1-\alpha_{2c}-(\sum(1-w_{cj})+1)/2}\exp\Big(-\frac{\beta_{2c}}{\sigma_{2i}^2}-\frac{1}{2\sigma_{2i}^2}(\sum_{j=1}^{G}(1-w_{cj})y_{ij}^2+\frac{\mu_{2c}^2}{r_{2c}})\Big) \\
&\quad\times\prod_{i=1}^{n_c}\exp\left\{-\frac{1}{2\sigma_{1i}^2}\Big(\sum_{j=1}^{G}w_{cj}+\frac{1}{r_{1c}}\Big)\Big(\mu_{1i}-\frac{\sum w_{cj}y_{ij}+\mu_{1c}/r_{1c}}{\sum w_{cj}+1/r_{1c}}\Big)^2\right\} \\
&\quad\times\prod_{i=1}^{n_c}\exp\left\{-\frac{1}{2\sigma_{2i}^2}\Big(\sum_{j=1}^{G}(1-w_{cj})+\frac{1}{r_{2c}}\Big)\Big(\mu_{2i}-\frac{\sum(1-w_{cj})y_{ij}+\mu_{2c}/r_{2c}}{\sum(1-w_{cj})+1/r_{2c}}\Big)^2\right\} \\
&\quad\times\prod_{i=1}^{n_c}\exp\Big\{\frac{1}{2\sigma_{1i}^2}\frac{(\sum w_{cj}y_{ij}+\mu_{1c}/r_{1c})^2}{\sum w_{cj}+1/r_{1c}}+\frac{1}{2\sigma_{2i}^2}\frac{(\sum(1-w_{cj})y_{ij}+\mu_{2c}/r_{2c})^2}{\sum(1-w_{cj})+1/r_{2c}}\Big\}.
\end{aligned}
$$

By the above we see that

$$
\begin{aligned}
\mu_{1i}|\sigma_{1i}^2,\mathbf{y} &\sim \mathrm{N}\left\{\frac{\sum w_{cj}y_{ij}+\mu_{1c}/r_{1c}}{\sum w_{cj}+1/r_{1c}},\sigma_{1i}^2\Big(\sum_{j=1}^{G}w_{cj}+\frac{1}{r_{1c}}\Big)^{-1}\right\}, \qquad (9.5)\\
\mu_{2i}|\sigma_{2i}^2,\mathbf{y} &\sim \mathrm{N}\left\{\frac{\sum(1-w_{cj})y_{ij}+\mu_{2c}/r_{2c}}{\sum(1-w_{cj})+1/r_{2c}},\sigma_{2i}^2\Big(\sum_{j=1}^{G}(1-w_{cj})+\frac{1}{r_{2c}}\Big)^{-1}\right\},
\end{aligned}
$$

and so when we integrate out μ_{1i} and μ_{2i}, for $i=1,\ldots,n_c$, we obtain

$$
\begin{aligned}
&\int L(\mathbf{w}_c,\boldsymbol{\theta}_c)\prod_{i=1}^{n_c}f(\mu_{1i}|\sigma_{1i}^2)f(\sigma_{1i}^2)f(\mu_{2i}|\sigma_{2i}^2)f(\sigma_{2i}^2)d(\mu_{1i})d(\mu_{2i}) \\
&= (2\pi)^{-(G+1)n_c}(r_{1c}r_{2c})^{-n_c/2}\beta_{1c}^{n_c\alpha_{1c}}\Gamma(\alpha_{1c})^{-n_c}\beta_{2c}^{n_c\alpha_{2c}}\Gamma(\alpha_{2c})^{-n_c} \\
&\quad\times\prod_{i=1}^{n_c}(\sigma_{1i}^2)^{-1-\alpha_{1c}-(\sum w_{cj}+1)/2}\exp\Big(-\frac{\beta_{1c}}{\sigma_{1i}^2}-\frac{1}{2\sigma_{1i}^2}(\sum_{j=1}^{G}w_{cj}y_{ij}^2+\frac{\mu_{1c}^2}{r_{1c}})\Big) \\
&\quad\times\prod_{i=1}^{n_c}(\sigma_{2i}^2)^{-1-\alpha_{2c}-(\sum(1-w_{cj})+1)/2}\exp\Big(-\frac{\beta_{2c}}{\sigma_{2i}^2}-\frac{1}{2\sigma_{2i}^2}(\sum_{j=1}^{G}(1-w_{cj})y_{ij}^2+\frac{\mu_{2c}^2}{r_{2c}})\Big) \\
&\quad\times\prod_{i=1}^{n_c}(2\pi)^{1/2}(\sigma_{1i}^2)^{1/2}\Big(\sum_{j=1}^{G}w_{cj}+\frac{1}{r_{1c}}\Big)^{-1/2} \\
&\quad\times\prod_{i=1}^{n_c}(2\pi)^{1/2}(\sigma_{2i}^2)^{1/2}\Big(\sum_{j=1}^{G}(1-w_{cj})+\frac{1}{r_{2c}}\Big)^{-1/2} \\
&\quad\times\prod_{i=1}^{n_c}\exp\Big\{\frac{1}{2\sigma_{1i}^2}\frac{(\sum w_{cj}y_{ij}+\mu_{1c}/r_{1c})^2}{\sum w_{cj}+1/r_{1c}}+\frac{1}{2\sigma_{2i}^2}\frac{(\sum(1-w_{cj})y_{ij}+\mu_{2c}/r_{2c})^2}{\sum(1-w_{cj})+1/r_{2c}}\Big\}
\end{aligned}
$$

$$
\begin{aligned}
= \;& (2\pi)^{-Gn_c}(r_{1c}r_{2c})^{-n_c/2}\beta_{1c}^{n_c\alpha_{1c}}\Gamma(\alpha_{1c})^{-n_c}\beta_{2c}^{n_c\alpha_{2c}}\Gamma(\alpha_{2c})^{-n_c} \\
& \times\Big(\sum_{j=1}^{G} w_{cj} + \frac{1}{r_{1c}}\Big)^{-n_c/2}\Big(\sum_{j=1}^{G}(1-w_{cj}) + \frac{1}{r_{2c}}\Big)^{-n_c/2} \\
& \times\prod_{i=1}^{n_c}(\sigma_{1i}^2)^{-1-\alpha_{1c}-\sum w_{cj}/2} \\
& \times\prod_{i=1}^{n_c}\exp\left(-\frac{1}{\sigma_{1i}^2}\Big\{\beta_{1c} + \frac{1}{2}(\sum_{j=1}^{G} w_{cj}y_{ij}^2 + \frac{\mu_{1c}^2}{r_{1c}}) - \frac{1}{2}\frac{(\sum w_{cj}y_{ij} + \mu_{1c}/r_{1c})^2}{\sum w_{cj} + 1/r_{1c}}\Big\}\right) \\
& \times\prod_{i=1}^{n_c}(\sigma_{2i}^2)^{-1-\alpha_{2c}-\sum(1-w_{cj})/2} \\
& \times\prod_{i=1}^{n_c}\exp\Big(-\frac{1}{\sigma_{2i}^2}\Big\{\beta_{2c} + \frac{1}{2}(\sum_{j=1}^{G}(1-w_{cj})y_{ij}^2 + \frac{\mu_{2c}^2}{r_{2c}}) \\
& \qquad\qquad -\frac{1}{2}\frac{(\sum(1-w_{cj})y_{ij} + \mu_{2c}/r_{2c})^2}{\sum(1-w_{cj}) + 1/r_{2c}}\Big\}\Big)
\end{aligned}
$$

From the above we see that

$$
\begin{aligned}
\sigma_{1i}^2|\mathbf{y} &\sim \text{IG}(\alpha_{1c}^*, \beta_{1c}^*), \\
\sigma_{2i}^2|\mathbf{y} &\sim \text{IG}(\alpha_{2c}^*, \beta_{2c}^*),
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha_{1c}^* &= \alpha_{1c} + \frac{1}{2}\sum_{j=1}^{G} w_{cj}, \\
\alpha_{2c}^* &= \alpha_{2c} + \frac{1}{2}\sum_{j=1}^{G}(1-w_{cj}),
\end{aligned}
$$

and

$$
\begin{aligned}
\beta_{1c}^* &= \Big\{\beta_{1c} + \frac{1}{2}(\sum_{j=1}^{G} w_{cj}y_{ij}^2 + \frac{\mu_{1c}^2}{r_{1c}}) - \frac{1}{2}\frac{(\sum w_{cj}y_{ij} + \mu_{1c}/r_{1c})^2}{\sum w_{cj} + 1/r_{1c}}\Big\}, \\
\beta_{2c}^* &= \Big\{\beta_{2c} + \frac{1}{2}(\sum_{j=1}^{G}(1-w_{cj})y_{ij}^2 + \frac{\mu_{2c}^2}{r_{2c}}) - \frac{1}{2}\frac{(\sum(1-w_{cj})y_{ij} + \mu_{2c}/r_{2c})^2}{\sum(1-w_{cj}) + 1/r_{2c}}\Big\}.
\end{aligned}
$$

When we integrate out σ_{1i}^2 and σ_{2i}^2, for $i = 1, \ldots, n_c$, we get

$$\begin{aligned}
& \int L(\mathbf{w}_c, \boldsymbol{\theta}_c) \prod_{i=1}^{n_c} f(\mu_{1i}|\sigma_{1i}^2) f(\sigma_{1i}^2) f(\mu_{2i}|\sigma_{2i}^2) f(\sigma_{2i}^2) d(\mu_{1i}) d(\mu_{2i}) d(\sigma_{1i}^2) d(\sigma_{2i}^2) \quad (9.6) \\
= \; & (2\pi)^{-Gn_c} (r_{1c} r_{2c})^{-n_c/2} \beta_{1c}^{n_c \alpha_{1c}} \Gamma(\alpha_{1c})^{-n_c} \beta_{2c}^{n_c \alpha_{2c}} \Gamma(\alpha_{2c})^{-n_c} \\
& \times \Big(\sum_{j=1}^{G} w_{cj} + \frac{1}{r_{1c}} \Big)^{-n_c/2} \Big(\sum_{j=1}^{G} (1 - w_{cj}) + \frac{1}{r_{2c}} \Big)^{-n_c/2} \\
& \times \Gamma(\alpha_{1c}^*)^{n_c} \Gamma(\alpha_{2c}^*)^{n_c} \prod_{i=1}^{n_c} (\beta_{1c}^*)^{-\alpha_{1c}^*} (\beta_{2c}^*)^{-\alpha_{2c}^*}
\end{aligned}$$

This likelihood is highly intractable. If we were to put any reasonable prior on $\mathbf{w}_c$ it seems unlikely that we could obtain a closed form solution of the marginal likelihood $L(\boldsymbol{\psi}_c)$. If we regard $\mathbf{w}_c$ as fixed maximizing the above marginal likelihood with respect to $\mathbf{w}_c$ and $\boldsymbol{\psi}_c$ also seems like a daunting task. However, we can approximate the above integral by plugging approximate Bayesian estimates of $(\mu_{1i}, \sigma_{1i}^2, \mu_{2i}, \sigma_{2i}^2)$ into the likelihood $L(\mathbf{w}_c, \boldsymbol{\theta}_c)$. This approximation only depends on $\mathbf{w}_c$ and not on any of the hyper-parameters and maximizing with respect to $\mathbf{w}_c$ turns out to be equivalent to the classification ML approach, discussed in section 7.7. We discuss this in detail in the following subsection.

9.2 Practical implementation

From (9.5) we get the posterior means of μ_{1i} and μ_{2i}

$$\begin{aligned}
\text{mean}(\mu_{1i}|\sigma_{1i}^2, \mathbf{y}) &= \frac{\sum_j w_{cj} y_{ij} + \mu_{1c}/r_{1c}}{\sum_j w_{cj} + 1/r_{1c}} \\
&= \tau_{1c} \mu_{1c} + (1 - \tau_{1c}) \frac{\sum_j w_{cj} y_{ij}}{\sum_j w_{cj}} \\
\text{mean}(\mu_{2i}|\sigma_{2i}^2, \mathbf{y}), &= \frac{\sum_j (1 - w_{cj}) y_{ij} + \mu_{2c}/r_{2c}}{\sum_j (1 - w_{cj}) + 1/r_{2c}} \\
&= \tau_{2c} \mu_{2c} + (1 - \tau_{2c}) \frac{\sum_j (1 - w_{cj}) y_{ij}}{\sum_j (1 - w_{cj})},
\end{aligned}$$

where $\tau_{1c} = (1/r_{1c})/(\sum_j w_{cj} + 1/r_{1c})$ and $\tau_{2c} = (1/r_{2c})/(\sum_j (1 - w_{cj}) + 1/r_{2c})$. We see that these posterior means shrink the classical weighted sample averages towards the prior means, μ_{1c} and μ_{2c}. Note that as the number of genes, G, increases the sums $\sum_j w_{cj}$ and $\sum_j (1 - w_{cj})$ increase and thus τ_{1c} and τ_{2c} decrease towards zero. Hence, we arrive at approximate estimates of the posterior means above

$$\hat{\mu}_{1i} = \frac{\sum_j w_{cj} y_{ij}}{\sum_j w_{cj}} \tag{9.7}$$

$$\hat{\mu}_{2i} = \frac{\sum_j (1 - w_{cj}) y_{ij}}{\sum_j (1 - w_{cj})} \tag{9.8}$$

From (9.4) we can we see that the conditional posteriors of σ^2_{1i} and σ^2_{2i} given μ_{1i} and μ_{2i} are given by

$$\begin{aligned}
\sigma^2_{1i} | \mu_{1i}, \mathbf{y} &\sim \text{IG}\Big(\alpha_{1c} + \frac{1}{2}\sum_{j=1}^{G} w_{cj}, \beta^{**}_{1c}\Big), \\
\sigma^2_{2i} | \mu_{2i}, \mathbf{y} &\sim \text{IG}\Big(\alpha_{2c} + \frac{1}{2}\sum_{j=1}^{G} (1 - w_{cj}), \beta^{**}_{2c}\Big),
\end{aligned}$$

where

$$\begin{aligned}
\beta^{**}_{1c} &= \beta_{1c} + \frac{1}{2}\Big\{ \sum_{j=1}^{G} w_{cj}(y_{ij} - \mu_{1i})^2 + \frac{1}{r_{1c}}(\mu_{1i} - \mu_{1c})^2 \Big\}, \\
\beta^{**}_{2c} &= \beta_{2c} + \frac{1}{2}\Big\{ \sum_{j=1}^{G} (1 - w_{cj})(y_{ij} - \mu_{2i})^2 + \frac{1}{r_{2c}}(\mu_{2i} - \mu_{2c})^2 \Big\}.
\end{aligned}$$

This gives us the posterior modes of σ^2_{1i} and σ^2_{2i}

$$\begin{aligned}
&\text{mode}(\sigma^2_{1i} | \mu_{1i}, \mathbf{y}) \\
= &\frac{\beta_{1c} + \frac{1}{2}\Big\{ \sum_j w_{cj}(y_{ij} - \mu_{1i})^2 + \frac{1}{r_{1c}}(\mu_{1i} - \mu_{1c})^2 \Big\}}{\alpha_{1c} + \sum_j w_{cj}/2 + 1} \\
= &\eta_{1c}\frac{\beta_{1c}}{\alpha_{1c} + 1} + (1 - \eta_{1c})\frac{\sum_j w_{cj}(y_{ij} - \mu_{1i})^2}{\sum_j w_{cj}} + \frac{(\mu_{1i} - \mu_{1c})^2}{2r_{1c}(\alpha_{1c} + \sum_j w_{cj}/2 + 1)},
\end{aligned}$$

and

$$\begin{aligned}
&\text{mode}(\sigma_{2i}^2|\mu_{2i},\mathbf{y}) \\
= &\frac{\beta_{2c}+\frac{1}{2}\Big\{\sum_j(1-w_{cj})(y_{ij}-\mu_{2i})^2+\frac{1}{r_{2c}}(\mu_{2i}-\mu_{2c})^2\Big\}}{\alpha_{2c}+\sum_j(1-w_{cj})/2+1} \\
= &\eta_{2c}\frac{\beta_{2c}}{\alpha_{2c}+1}+(1-\eta_{2c})\frac{\sum_j(1-w_{cj})(y_{ij}-\mu_{2i})^2}{\sum_j(1-w_{cj})}+\frac{(\mu_{2i}-\mu_{2c})^2}{2r_{2c}(\alpha_{2c}+\sum_j(1-w_{cj})/2+1)},
\end{aligned}$$

where $\eta_{1c}=(\alpha_{1c}+1)/(\alpha_{1c}+\sum_j w_{cj}/2+1)$ and $\eta_{2c}=(\alpha_{2c}+1)/(\alpha_{2c}+\sum_j(1-w_{cj})/2+1)$. If we plug in $\hat{\mu}_{1i}$ and $\hat{\mu}_{2i}$ we see that the posterior modes are of shrinkage form as well, where the weighted sample variances are being shrunk towards the prior modes of σ_{1i}^2 and σ_{2i}^2. However, there are additional terms

$$\frac{(\hat{\mu}_{1i}-\mu_{1c})^2}{2r_{1c}(\alpha_{1c}+\sum_j w_{cj}/2+1)}, \quad \text{and} \quad \frac{(\hat{\mu}_{2i}-\mu_{2c})^2}{2r_{2c}(\alpha_{2c}+\sum_j(1-w_{cj})/2+1)},$$

but they tend towards zero as the number of genes, G, increases. The shrinkage factors η_{1c} and η_{2c} also tend to zero as G increases and so we arrive at approximate estimates of the posterior modes

$$\hat{\sigma}_{1i}^2 = \frac{\sum_j w_{cj}(y_{ij}-\hat{\mu}_{1i})^2}{\sum_j w_{cj}}, \tag{9.9}$$

$$\hat{\sigma}_{2i}^2 = \frac{\sum_j(1-w_{cj})(y_{ij}-\hat{\mu}_{2i})^2}{\sum_j(1-w_{cj})}. \tag{9.10}$$

Now instead of maximizing the intractable marginal likelihood in (9.6) we can plug these approximate Bayes estimates of μ_{1i}, μ_{2i}, σ_{1i}^2 and σ_{2i}^2 into the likelihood $L(\mathbf{w}_c,\boldsymbol{\theta}_c)$, given in (9.3), and maximize with respect to w_{cj}. But recall that the maximization of the likelihood $L(\mathbf{w}_c,\boldsymbol{\theta}_c)$, under the assumption that $\mathbf{w}_c$ is fixed was discussed in section 7.7 and involved iterating between the following two steps

1. For a fixed w_{cj} update the parameters according to

$$\begin{aligned}
\hat{\mu}_{1i} &= \frac{\sum_j w_{cj}y_{ij}}{\sum_j w_{cj}} \\
\hat{\mu}_{2i} &= \frac{\sum_j(1-w_{cj})y_{ij}}{\sum_j(1-w_{cj})}
\end{aligned}$$

and

$$\begin{aligned}\hat{\sigma}_{1i}^2 &= \frac{\sum_j w_{cj}(y_{ij}-\hat{\mu}_{1i})^2}{\sum_j w_{cj}},\\ \hat{\sigma}_{2i}^2 &= \frac{\sum_j (1-w_{cj})(y_{ij}-\hat{\mu}_{2i})^2}{\sum_j (1-w_{cj})}.\end{aligned}$$

2. For current values of the parameters, $(\hat{\mu}_{1i}, \hat{\sigma}_{1i}^2, \hat{\mu}_{1i}, \hat{\sigma}_{2i}^2)$ update w_{cj} according to whether

$$\sum_{i=1}^{n_c} \log \phi(y_{ij}|\hat{\mu}_{1i}, \hat{\sigma}_{1i}^2) > \sum_{i=1}^{n_c} \log \phi(y_{ij}|\hat{\mu}_{2i}, \hat{\sigma}_{2i}^2)$$

Note that step 1. is exactly the Bayesian approximate estimates of the parameters, from above. Thus the maximization in section 7.7 is in reality an approximate solution to the above more realistic model.

CHAPTER 10

ANALYSIS

In this chapter we analyze the Erasmus data set. The data was collected at Erasmus University Medical Center (Rotterdam) between 1990-2008 and involves methylation and expression profiles, across the whole genomes, of $n = 344$ patients with Acute Myeloid Leukemia (AML). As expression data has become increasingly familiar to Statisticians over the last few years we will only describe the methylation data in detail in section 10.1. However, if we abstract away from this application, we can think of both methylation and expression as ON/OFF processes where a methylated/expressed DNA fragment is considered ON and a non-methylated/non-expressed fragment is considered OFF. In section 10.2 we provide a detailed clustering analysis of the Erasmus data and in section 10.3 we conduct a brief discriminant analysis on a well characterized subset of the data. We end this chapter with a section on how to identify the discriminating genes of a given partition.

10.1 Data description

In this section we describe the Erasmus methylation data set in detail. For a more detailed biological description see Figueroa et al. (2010). Briefly, one μg of gDNA is digested with two enzymes, HpaII and its isoschizomer MspI, separately. The products of these two digestions are then PCR amplified, which results in $G = 25,626$ pairs of DNA fragments, ranging from 200-2000bp in size. This corresponds to roughly $14,000$ unique genes. For each patient the pairs of DNA fragments are then hybridized onto the probes of a microarray. By co-hybridizing and comparing signal intensities from the MspI (representing the entire genome), and HpaII (hypomethylated) fractions, the methylation levels at each promoter or genomic locus can be determined. At each of the $G = 25,626$

probes the signal intensities are compared with the log signal ratio. More precisely, for patient i the measurement at each probe, j, is defined as $y_{ij} = \log(\text{HpaII}_{ij}/\text{MspI}_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, G$. The HpaII signal is only high at non-methylated regions of the genome, whereas MspI gives high signal at both methylated and non-methylated regions. A low value of the log-ratio thus indicates methylation and a high value non-methylation. However, these values do not explicitly tell us whether a DNA fragment is methylated or not. If we look at a histogram of the log-ratios for each patient i we see roughly a bimodal distribution, see Figure 10.1, where the left mode corresponds to methylated genes and the right mode to non-methylated genes. Genes with log-ratios far to the left are very likely to be methylated and genes far to the right are very likely to be non-methylated. But, for genes located roughly in between the two modes it is not clear whether they are methylated or not.

Recall that in the model based clustering algorithm we only looked at a subset of genes, $J_d \subset \{1, \ldots, G\}$, that we declared discriminating. For those genes we assumed that in a given cluster, c, they either methylated for all patients in that cluster or they did not methylate for all patients in the cluster. This model assumption is presented graphically in Figure 10.2. We introduced in (7.1) the latent cluster specific methylation indicator, w_{cj}, and assumed for each cluster, c, that $w_{cj} \sim Bernoulli(\pi_{1c})$, independently across $j \in J_d$. This lead to the marginal distribution of $\mathbf{y}_c = (\mathbf{y}_i)_{i \in c}$ given in (7.6). The marginal distribution of the data for patient i in cluster c is given by

$$f(\mathbf{y}_i) = \prod_{j \in J_d} \left\{ \pi_{1c} \phi(y_{ij} | \mu_{1i}, \sigma^2_{1i}) + \pi_{0c} \phi(y_{ij} | \mu_{2i}, \sigma^2_{2i}) \right\}, \tag{10.1}$$

and note that the individual densities for patients $i \in c$ share one parameter in common, π_{1c}, but are allowed to have individual specific means and variances. In Figure 10.2 we can see histograms of four different patients. Patients number 16 and 18 share a cluster but do clearly not have the same mixture distribution. This is what we call the array effect. Since each patient has data measured on physically different microarrays, which

introduces variability across patients, we do not assume identical mixture distributions within clusters, but only require that each gene methylates identically within the cluster. This is fundamentally different from the usual model based clustering setup where we assume purely cluster specific distributions. By introducing these individual specific parameters, within clusters, we allow for more flexibility in the modeling of the data and in the process we account for the array effect. Also, note that even though we are introducing 4 extra parameters for each patient in the study we have several thousand measurements per patient to estimate them. One of the most important features of our model based clustering algorithm is that we not only cluster the patients, but also obtain information about the cluster specific gene methylations. We make predictions about whether or not a given gene, j, methylates in cluster c based on the posterior expectations of the latent methylation indicators. If we were to make predictions about w_{cj} based only on the data for a single patient i (see (10.1)) we would look at the posterior expectation

$$E[w_{cj}|\mathbf{y}_i] = \frac{\pi_{1c}\phi\big(y_{ij}|\mu_{1i},\sigma^2_{1i}\big)}{\pi_{1c}\phi\big(y_{ij}|\mu_{1i},\sigma^2_{1i}\big)+\pi_{0c}\phi\big(y_{ij}|\mu_{2i},\sigma^2_{2i}\big)}. \tag{10.2}$$

The closer this ratio is to 1 the more likely it is that gene j is methylated in cluster c and the closer it is to 0 the more likely it is that it is hypo-methylated. As we mentioned above, if we only look at the methylation profile of a single individual, see Figure 10.1, many of the genes fall roughly in between the two modes of the mixture distribution and reliable methylation predictions cannot be made for these genes based on the individual profile alone. But since we are assuming identical methylation profiles within clusters we actually borrow strength in methylation prediction across all patients $i \in c$. This can be readily seen by the formula for the posterior expectation of the methylation indicator given all the data, in cluster c

$$E[w_{cj}|\mathbf{y}_c] = \frac{\pi_{1c}\prod_{i\in c}\phi\big(y_{ij}|\mu_{1i},\sigma^2_{1i}\big)}{\pi_{1c}\prod_{i\in c}\phi\big(y_{ij}|\mu_{1i},\sigma^2_{1i}\big)+\pi_{0c}\prod_{i\in c}\phi\big(y_{ij}|\mu_{2i},\sigma^2_{2i}\big)}.$$

Compare this to the formula provided in (10.2). Thus, even though we are not able to make accurate predictions about the methylation status of a given gene, j, for any

of the individuals $i \in c$ separately, it is still possible that we can predict that $w_{cj} = 1$ (or $w_{cj} = 0$) with high estimated posterior probability. In fact most of the gene methylation indicators have estimated posterior expectations close to zero or one. This is not surprising given the asymptotic result of Theorem 7.5.1, where we showed that as the number of patients within a cluster grows we can more accurately predict methylation.

10.2 Clustering results

For the Erasmus patients we have methylation data on $G_1 = 25,626$ DNA fragments and expression data on $G_2 = 54,675$ DNA fragments. Before running the hierarchical clustering algorithm we determined suitable subsets of the two data sets. For the methy-

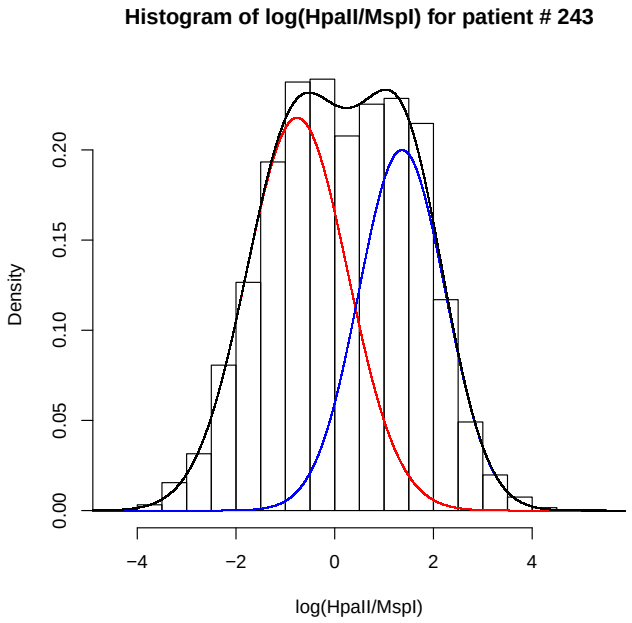

Figure 10.1: A histogram of the log signal ratio, $\log(\text{HpaII/MspI})$, for patient number 234, along with a two component Gaussian mixture fit. The left mode (red density) corresponds to methylated genes and the right mode (blue density) to non-methylated genes. The black density represents the mixture density of the two normals.

lation and expression data we declared discriminating all DNA fragments with standard deviations > 1 and > 0.5 respectively. This resulted in subsets $J_d^1 \subset \{1, \ldots, G_1\}$ and $J_d^2 \subset \{1, \ldots, G_2\}$ with $\#J_d^1 = 3,745$ and $\#J_d^2 = 3,370$. In the next three subsections we discuss the results from the hierarchical clustering algorithm applied to J_d^1, J_d^2, and $J_d^1 \cup J_d^2$. In these subsections we also discuss the effect of applying the two way CEM algorithm of section 7.7 to the results of the hierarchical algorithm. In subsection 10.2.4 we discuss the partitions obtained by applying the full data two way CEM of section 8.4 to the results of the hierarchical algorithm. Finally, in subsection 10.2.5 we give a detailed sensitivity and specificity analysis of the different clustering results in terms of three known and well characterized subtypes.

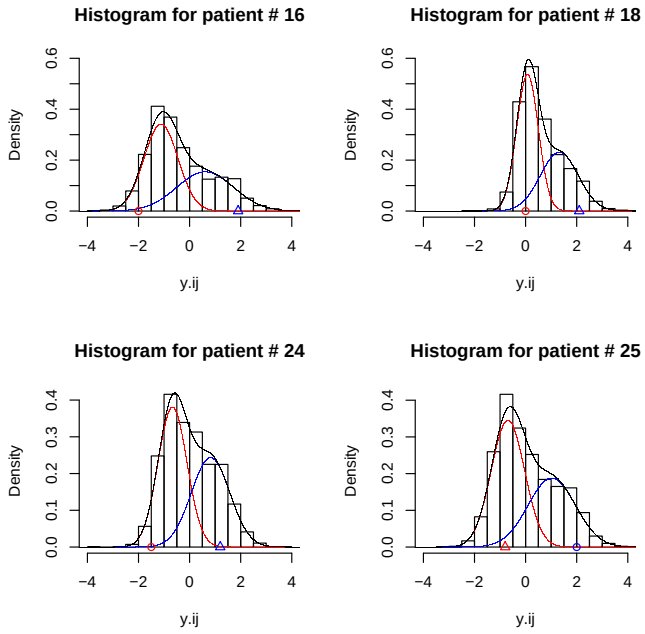

Figure 10.2: A histogram of $(y_{ij})_{j \in J_d}$ for 4 different patients and the fits obtained by fitting the model of chapter 7. Patients 16 and 18 are members of the same cluster, whereas patients 24 and 25 are not. The marks (triangle and bullet) represent the positions of two distinct genes and the color determines whether or not those genes were methylated (red methylation, blue non-methylation). For patients in the same cluster (16 and 18) we expect to see agreement in methylation as seen above, but for patients in different clusters (24 and 25) we expect to see more of a disagreement.

10.2.1 Clustering based on methylation data

We first ran the hierarchical clustering algorithm of section 7.3 on the methylation data, J_d^1. This resulted in a total of 344 candidate partitions, call them $M_1, \ldots, M_{344}$, with numbers of clusters taking on the values $K = 1, \ldots, 344$, respectively. If we plot the log-likelihood values of these partitions against the numbers of clusters we see a curve that has the lowest value at $K = 344$, rises towards smaller numbers of clusters, peaks at $K = 17$ and then drops again down to $K = 1$. In Figure 10.3 we can see a portion of this plot around the peak (see lower curve marked with "o"). As the log-likelihood curve peaks at $K = 17$ clusters the corresponding partition is the desired partition. However, one of the clusters of that partition involves a single patient and at the next merge that patient is put into a bigger cluster. Since the log-likelihood values do not differ that much between the partition with 16 clusters and the one with 17 clusters we could argue that the singleton free 16 cluster partition is a more reasonable clustering result. Figueroa et al. (2010) also used J_d^1 as the set of discriminating genes and arrived at a partition with 16 clusters as well, but the cutoff of $K = 16$ was selected heuristically. We chose the partition corresponding to $K = 16$ to get a direct comparison of the two methods. In Figure 10.4 we see a graphical comparison of our clustering result and that of Figueroa et al. (2010). The graph displays a correlation heat map of the 344 patients. The first diagonal strip corresponds to the correlation based clustering result of Figueroa et al. (2010) and the second strip corresponds to our likelihood based result. Clusters number 1, 3, and 6, on the first diagonal strip, correspond to three known and well characterized clusters, inv(16), $t(8;21)$ and $t(15;17)$, respectively. Both the correlation based hierarchical clustering method and the likelihood based method do a good job at picking up these three robust clusters. For a detailed sensitivity and specificity analysis of the clustering results, in terms of these three well characterized subtypes of AML, see section 10.2.5. According to Figueroa et al. (2010) each of the eight clusters number

4, 9, 10, 11, 12, 13, 14, and 16 had some biological features common to the cluster members. Again, by looking at the graph we can see a substantial agreement between the two methods in determining these clusters. However, methylation clusters 2, 5, 7, 8, and 15 were solely defined by their DNA methylation profiles and could not be explained by other biological features of cluster members. For three of these novel clusters, clusters number 7, 8, and 15, we do not see much of an overlap between the two methods. However, given that these are ill-defined clusters it is hard to make a statement about the validity of one method over the other in terms of these discrepancies. To improve upon our desired partition we used the CEM algorithm described in section 7.7. For each $K = 11, \ldots, 30$ (the values around the peak of Figure 10.3) we ran the CEM algorithm using $\mathcal{C}_K$, from the hierarchical clustering algorithm, as the initial partition. The resulting partitions led to an improvement in the log-likelihoods for each K. The curve marked with "$\times$" in Figure 10.3 shows the log-likelihood values of these partitions plotted against the numbers of clusters. This curve peaks at $K = 18$ and has second and third highest log-likelihood values at $K = 17$ and $K = 16$ respectively. Since each of the two partitions with $K = 18$ and $K = 17$ have single patient clusters, but the one with $K = 16$ is singleton free, we would argue in favor of using the 16 cluster partition. When we ran the CEM algorithm on $\mathcal{C}_{16}$ the algorithm converged in two steps. It reallocated exactly 13 patients in the first step of the algorithm but had then converged. We examined the reallocation posterior probabilities of the CEM algorithm, $P(X_{ic} = 1|\text{Data})$, for each patient i and cluster c, and saw that 343 of the 344 patients had posterior probabilities approximately equal to 1 for one cluster and 0 for the other ones. Only one patient had a posterior probability less than 1, for the most likely cluster, but the value was 0.88 which is still high. We thus strongly recommend using the CEM algorithm rather than the EM algorithm as the results should not be affected dramatically and the CEM algorithm is easier to implement and faster to run.

10.2.2 Clustering based on expression data

We also ran the hierarchical clustering algorithm of section 7.3 on the expression data, J_d^2. This resulted in the candidate partitions $E_1, \ldots, E_{344}$ with cluster sizes $K = 1, \ldots, 344$, respectively. The partition corresponding to the maximum likelihood had 17 clusters, see curve marked "o" in Figure 10.5. We then ran the CEM algorithm, using each of the partitions, E_K, as the initial partition. This lead to an improvement in likelihood for each cluster size, K. The curve marked "x" in Figure 10.5 shows a part of the log-likelihood curve for the partitions obtained through CEM. Note that the maximum likelihood is now obtained at the partition corresponding to $K = 15$, but not $K = 17$. Since there was not a vast difference in these log-likelihood values we used

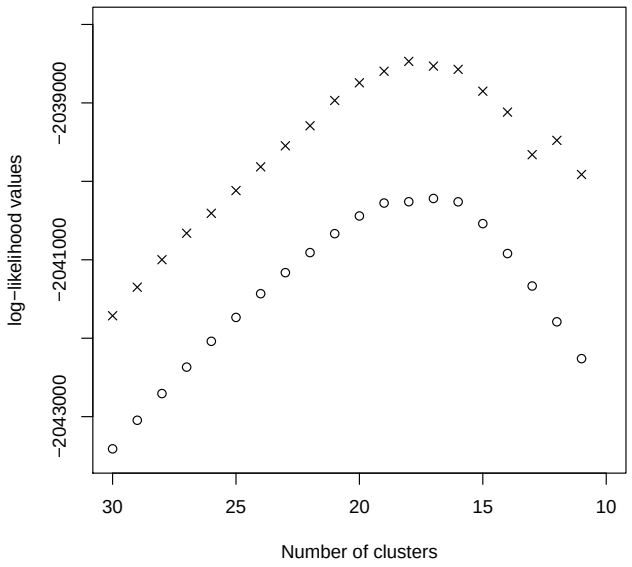

Figure 10.3: From the hierarchical clustering algorithm, applied to the methylation data, we obtain candidate partitions $M_1, \ldots, M_{344}$ with numbers of clusters $K = 1, \ldots, 344$ respectively. The curve marked with "o" shows the log-likelihood values of $M_{30}, M_{29}, \ldots, M_{11}$ plotted against the numbers of clusters. The curve marked with "$\times$" shows the log-likelihood curve of the partitions obtained by running the CEM algorithm, using each of the partitions, M_K, as initial partitions.

$K = 17$ in the analysis that follows to get a direct comparison of the different clustering results based on the expression data. The likelihood based method applied to the expression data alone resulted in a reasonable partition. In particular it did a good job at picking up the three robust clusters, inv(16), $t(15; 17)$ and $t(8; 21)$, just like when using the methylation data. The expression data clustering did a better job, than the methylation data clustering, at picking up the clusters $t(15; 17)$ and $t(8; 21)$, but a worse job at picking up cluster inv(16); see sensitivity and specificity analysis subsection below.

10.2.3 Clustering based on both data types

We finally ran the multiple platform hierarchical clustering algorithm on both methylation and expression data, J_d^1 and J_d^2, simultaneously. We can see a plot of the log-likelihood values against the numbers of clusters in Figure 10.6 (the curve marked with "o"). Then we ran the two way CEM algorithm of section 7.7, using these partitions as initial partitions, and we can see a plot of the resulting log-likelihood values in Figure

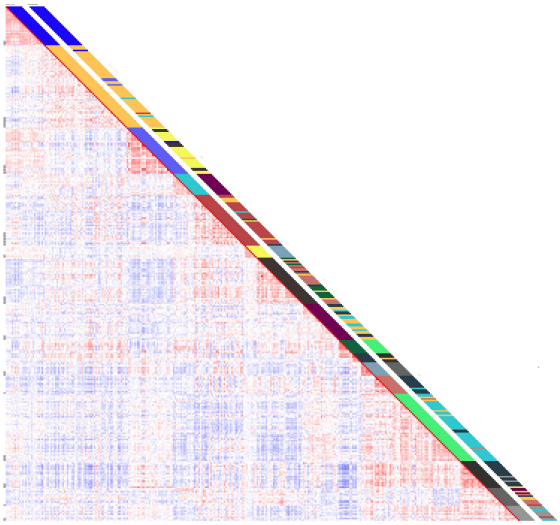

Figure 10.4: The above Figure shows a correlation heat map for the 344 AML patients. The first diagonal strip represents the clustering result of Figueroa et al. (2010) and the second diagonal strip corresponds to our likelihood based clustering result.

10.6 as well, curve marked with "×". For both curves the likelihood is maximized at $K = 14$. The multiple platform method applied to both data types simultaneously did a better job at picking up the three robust clusters, inv(16), $t(15; 17)$ and $t(8; 21)$, than the single platform method applied to each of the two data types separately. There were some patients that were misclassified (in terms of the three robust clusters) using the expression data but not misclassified using the methylation data and vice versa. The joint analysis for the most part corrected for these types of one sided mistakes, see sensitivity and specificity analysis in subsection 10.2.5.

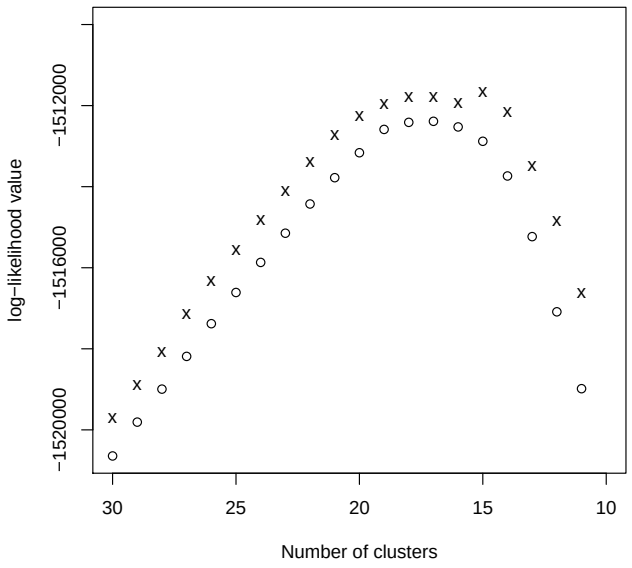

Figure 10.5: From the hierarchical clustering algorithm, applied to the expression data, we obtain candidate partitions $E_1, \ldots, E_{344}$ with numbers of clusters $K = 1, \ldots, 344$ respectively. The curve marked with "o" shows the log-likelihood values of $E_{30}, E_{29}, \ldots, E_{11}$ plotted against the numbers of clusters. The curve marked with "x" shows the log-likelihood curve of the partitions obtained by running the CEM algorithm, using each of the partitions, E_K, as initial partitions.

10.2.4 Two way CEM applied to all data

We also applied the two way CEM algorithm of section 8.4 to the full data sets using all $G_1 = 25,626$ and $G_2 = 54,675$ DNA fragments from the methylation and expression data respectively. In terms of classifying the three robust clusters this gave an identical result to the partial data analysis using J_d^1 for the methylation data but did worse using J_d^2 for the expression data. We also got a less favorable result applying CEM to all DNA fragments of both data types than only to $J_d^1 \cup J_d^2$. We will discuss the sensitivity and specificity of these results along with the results of sections 10.2.1-10.2.3 in more detail in the following section.

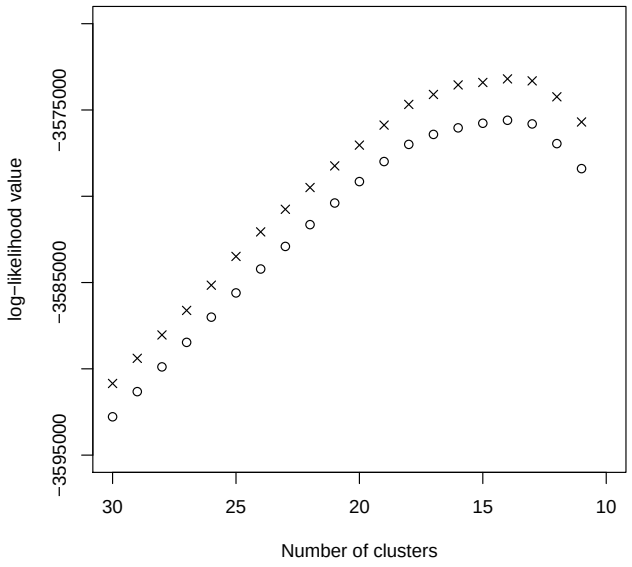

Figure 10.6: From the hierarchical clustering algorithm, applied to both the methylation and the expression data, we obtain candidate partitions $ME_1, \ldots, ME_{344}$ with numbers of clusters $K = 1, \ldots, 344$ respectively. The curve marked with "o" shows the log-likelihood values of $ME_{30}, ME_{29}, \ldots, ME_{11}$ plotted against the numbers of clusters. The curve marked with "$\times$" shows the log-likelihood curve of the partitions obtained by running the CEM algorithm, using each of the partitions, ME_K, as initial partitions.

10.2.5 Sensitivity and specificity analysis

Among the $n = 344$ patients there were a total of 62 patients with three known and well characterized subtypes of AML [sample sizes in brackets], inv(16) [$n_1 = 28$], $t(15;17)$ [$n_2 = 10$] and $t(8;21)$ [$n_3 = 24$]. In this section we summarize for each of these three subtypes the sensitivity and specificity of our clustering results. Before presenting our numerical results let us first recall the definitions of sensitivity and specificity for any given subtype, S. Sensitivity is the percentage of patients with subtype S who are identified as having it

$$\text{sensitivity}(S) = \frac{\#\text{true positives}(S)}{\#\text{true positives}(S) + \#\text{false negatives}(S)}.$$

Specificity is the percentage of patients not with subtype S who are identified as not having it

$$\text{specificity}(S) = \frac{\#\text{true negatives}(S)}{\#\text{true negatives}(S) + \#\text{false positives}(S)},$$

In the above "positives(S)" are those patients who are declared to have subtype S and "negatives(S)" are those who are declared not to have it. Note however, that in cluster analysis we only cluster the patients into different groups but make no statements about which group corresponds to which subtype. Thus the above definitions are not well defined in cluster analysis. A straight forward remedy is to look at the clustering labels of all patients with a certain known subtype, S, and identify the most frequent clustering label. The cluster corresponding to that label will be declared as the "postives(S)" and all other patients will be declared as the "negatives(S)".

In Table 10.1 we summarize the sensitivity and specificity of the clustering results based on the discriminating gene sets J_d^1 and J_d^2 described in sections 10.2.1-10.2.3. We let M, E, and ME denote the partitions obtained through the use of the hierarchical algorithm on the methylation data, expression data, and both data types respectively. We

Table 10.1: The sensitivity and specificity of the clustering results based on the methylation data, J_d^1, the expression data, J_d^2, and both data types, $J_d^1 \cup J_d^2$ for the three robust clusters, inv$(16)[n_1 = 28]$, $t(15;17)[n_2 = 10]$ and $t(8;21)[n_3 = 24]$. The sensitivity and specificity of the correlation clustering result (based on J_d^1) is provided for comparison.

Sensitivity (# false negatives in parentheses)

subtype	COR	M	E	ME	M+CEM	E+CEM	ME+CEM
inv(16)	0.929(2)	0.964(1)	0.857(4)	0.964(1)	0.964(1)	0.857(4)	0.964(1)
$t(15;17)$	0.800(2)	0.800(2)	1.000(0)	1.000(0)	0.800(2)	1.000(0)	1.000(0)
$t(8;21)$	0.917(2)	0.875(3)	0.917(2)	0.958(1)	0.875(3)	0.917(2)	0.958(1)

Specificity (# of false positives in parentheses)

subtype	COR	M	E	ME	M+CEM	E+CEM	ME+CEM
inv(16)	1.000(0)	0.997(1)	0.997(1)	0.997(1)	0.997(1)	0.997(1)	1.000(0)
$t(15;17)$	1.000(0)	1.000(0)	1.000(0)	1.000(0)	1.000(0)	1.000(0)	1.000(0)
$t(8;21)$	0.972(9)	0.994(2)	1.000(0)	0.991(3)	0.994(2)	1.000(0)	0.994(2)

let M+CEM, E+CEM, ME+CEM denote the partitions obtained by running the two way CEM algorithm of section 7.7 on J_d^1, J_d^2 and $J_d^1 \cup J_d^2$ (using M, E and ME as the initial partitions) respectively. The sensitivity and specificity of the correlation clustering result for the methylation data, denoted COR, is provided as well for comparison. Comparing the COR result and the M result we see that M is more sensitive, in terms of subtype inv(16), having 1 less false negative than COR. M is however less specific with one more false positive than COR. The two methods are directly comparable in terms of sensitivity and specificity of $t(15;17)$. For subtype $t(8;21)$ COR is more sensitive with one less false negative, but M is substantially more specific with 7 less false positives. This fairly big difference in specificity of $t(8;21)$ might suggest that the likelihood based method is a more stable method than the hierarchically based correlation clustering method. By comparing the columns M, E, ME to the columns M+CEM, E+CEM, ME+CEM we can see that applying the CEM algorithm (using M, E and ME as initial partitions) does not affect the sensitivities of the three clusters for any of the data, methylation, expression or both data types. Also, the specificity of M and E does not increase by applying the CEM algorithm. However, comparing columns ME and ME+CEM we see an improvement

Table 10.2: The sensitivity and specificity of the clustering results based on applying CEM to the candidate partitions of the hierarchical clustering algorithm. This Table uses all the methylation data, M_{full}, all the expression data E_{full} and all joint data ME_{full}.

Sensitivity (# of false negatives in parentheses)

subtype	M_{full}+CEM	E_{full}+CEM	ME_{full}+CEM
inv(16)	$0.964(1)$	$0.679(\mathbf{9})$	$0.929(2)$
$t(15;17)$	$0.800(2)$	$1.000(0)$	$1.000(0)$
$t(8;21)$	$0.875(3)$	$0.875(3)$	$0.958(1)$

Specificity (# of false positives in parentheses)

subtype	M_{full}+CEM	E_{full}+CEM	ME_{full}+CEM
inv(16)	$0.997(1)$	$0.968(\mathbf{10})$	$0.991(3)$
$t(15;17)$	$1.000(0)$	$1.000(0)$	$1.000(0)$
$t(8;21)$	$0.994(2)$	$1.000(0)$	$0.994(2)$

in specificity by running the CEM algorithm, ME+CEM having one less false positive for both inv(16) and $t(8;21)$. Another thing to note is that if we compare the ME+CEM column to M+CEM and E+CEM we see that the joint analysis usually does as well or better than each analysis separately both in terms of sensitivity and specificity. The only exception is that E+CEM is more specific than ME+CEM with 2 less false positives for subtype $t(8;21)$.

In Table 10.2 we have summarized the sensitivities and specificities of the three robust clusters using the results based on two way CEM applied to all $G_1 = 25,626$ and $G_2 = 54,675$ DNA fragments of the methylation and expression data, respectively. We let M_{full}+CEM, E_{full}+CEM and ME_{full}+CEM denote the partitions obtained by applying full data CEM to the partitions M, E, and ME respectively. The clustering based on the methylation data, M_{full}+CEM gives the same sensitivity and specificity results as M+CEM based on J_d^1 in Table 10.1. This suggests that we can run an analysis on the methylation data using all genes rather than selecting discriminating genes beforehand in an ad hoc way. However, the clustering result based on the full expression data is not as

favorable. It seems that by using all the data rather than J_d^2 (#$J_d^2 = 3,370$) we actually do worse. It is interesting to note though that the only problem seems to be the low sensitivity and specificity of subtype inv(16) (# false negatives=9 and # false positives = 10 for E_{full}+CEM). The expression data in fact gives better sensitivity and specificity, than the methylation data, for both of the other clusters, $t(15;17)$ and $t(8;21)$. Note also that the number of false negative $t(8;21)$'s went up from 2 for E+CEM to 3 for E_{full}+CEM. But if we look more carefully at the resulting partitions we see that the extra false negative is in fact falsely allocated to the problematic cluster inv(16). A potential explanation for the poorer performance of the two way CEM applied to the full expression data is that gene expression might simply not discriminate inv(16) very well from the other subtypes. Thus, using more expression data in the analysis could result in a more noisy result. Finally, we note that ME_{full}+CEM does as well as ME+CEM in terms of sensitivity and specificity of $t(15;17)$ and $t(8;21)$, but has one more false negative and two more false positive inv(16)'s than ME+CEM. Since we have roughly twice as many expression data, $G_2 = 54,675$, than methylation data, $G_1 = 25,626$, it is not unreasonable to suspect that the influences of the poor performance of the expression data carries over to the joint analysis.

10.2.6 Clustering the robust clusters only

The poor performance of the CEM algorithm applied to the full expression data in the previous subsection raises some questions. We decided to run a simple clustering analysis on the 62 patients with the three known subtypes, inv(16), $t(15;17)$ and $t(8;21)$ only to compare the full data CEM to the partial data (J_d^1 and J_d^2) CEM. We ran the hierarchical clustering algorithm on the 62 patients using the methylation data, J_d^1, expression data J_d^2 and both types, $J_d^1 \cup J_d^2$. This resulted in the partitions M, E and ME

which we then used as initial partitions for the CEM algorithms. For each of J_d^1, J_d^2 and $J_d^1 \cup J_d^2$ the hierarchical clustering algorithm arrived at a 3-cluster partition. For each data set, methylation, expression and both types, we then ran a partial data CEM and full data CEM using each of the three partitions M, E and ME as initial partitions. Table 10.3 shows the number of misclassified patients (out of 62) for the different CEMs and the different initial partitions. We let $\mathrm{M}_{J_d^1}^{CEM}$, $\mathrm{E}_{J_d^2}^{CEM}$, $\mathrm{ME}_{J_d^1 \cup J_d^2}^{CEM}$ denote the partitions obtained by applying the CEM algorithm to J_d^1, J_d^2 and $J_d^1 \cup J_d^2$, respectively. Similarly we let $\mathrm{M}_{\text{full}}^{CEM}$, $\mathrm{E}_{\text{full}}^{CEM}$, $\mathrm{ME}_{\text{full}}^{CEM}$ denote the partitions obtained by applying the CEM algorithm to the full methylation, expression and both data types, respectively. First let us note that the hierarchical clustering algorithm correctly identifies a 3-cluster partition of the 62 patients for J_d^1, J_d^2 and $J_d^1 \cup J_d^2$. By examining the column N of Table 10.3 we see that the partition M, based on the methylation data, misclassifies 3 patients, E, based on the expression data, misclassifies 1 and ME, based on both data types, perfectly classifies the 62 patients. Running the partial CEM and full CEM on the methylation data converges to the same partition, with 2 misclassifications, regardless of whether the initial partition is M, E, or ME. In the case when M is the initial partition there is an improvement in number of misclassifications, but when we start from E or ME we converge to a less accurate partition. Running the partial CEM and full CEM on the expression data leads to slightly different results. The partial CEM always converges to the same partition, with one misclassification, regardless of the initial partition. Running the full CEM using E as the initial partition converges to that same partition but when we use M and ME as the initial partitions we converge to the true partition. For the joint methylation and expression data analysis the partial CEM and full CEM gave identical results. The above suggests that using the full data leads to a better result than simply using the subsets J_d^1 and J_d^2.

Table 10.3: Number of misclassifications (out of 62) after applying partial and full CEMs to the different data types under different initial partitions. The column N denotes the numbers of misclassifications of the partitions M, E, and ME before applying the CEM algorithm.

Initial $\mathcal{C}$	N	$\text{M}_{J_d^1}^{CEM}$	$\text{M}_{\text{full}}^{CEM}$	$\text{E}_{J_d^2}^{CEM}$	$\text{E}_{\text{full}}^{CEM}$	$\text{ME}_{J_d^1 \cup J_d^2}^{CEM}$	$\text{ME}_{\text{full}}^{CEM}$
M	3	2	2	1	0	1	1
E	1	2	2	1	1	0	0
ME	0	2	2	1	0	0	0

10.2.7 Discussion about the clustering results

In the above subsections we seem to be arriving at contradicting conclusions. When we clustered the complete Erasmus data set with $n = 344$ patients the full expression data CEM lead to a worse result than the partial CEM. However, when we applied the partial and full CEM on the three robust clusters $(n = 62)$ only, the full data analysis did better. In the following we present a potential explanation for the above phenomena. Note that we get very poor clustering result if we use all the genes in the hierarchical clustering algorithm rather than a subset based on variance cutoff. This is due to the fact that there are several thousand genes that behave almost identical across all $n = 344$ patients. This lead us to the notion of discriminating genes and the gene importance indicator γ_j. It seems reasonable to assume that if we do a good job at finding these discriminating genes we should accurately reallocate patients to different clusters through the CEM algorithm. But there is a slight dilemma. If we have a partition with multiple clusters it is possible that two distinct clusters differ in methylation on say $1,000$ genes. These $1,000$ genes methylating differently across these two clusters would clearly be declared discriminating for the given partition. If we look at a different pair of clusters it is possible that they would not necessarily differ in methylation on the same set of $1,000$ genes. In fact they might differ in methylation on a completely different set of genes. All genes that differ in methylation for at least one pair of clusters are declared discriminating and so the full set of discriminating genes for the partition is always go-

ing to be bigger than any set of genes that discriminate between a single pair of clusters. The CEM algorithm involves reallocating a patient from one cluster to another. The cluster specific likelihood of a patient is evaluated at the current cluster and all other clusters. If the likelihood of a different cluster is greater than the one of the current cluster we reallocate the patient. But if two clusters only differ in methylation on say $1,000$ genes and the set of discriminating genes was declared to contain say $10,000$ genes then clearly we are adding a lot of noise to the comparison of likelihoods of these two clusters. When we apply the full data CEM on the three robust clusters the discriminating gene set of the expression data involves a total of $4,059$ genes. But when the full data CEM is applied to the complete Erasmus expression data set ($n = 344$) the set of declared discriminating genes involves a total of $9,856$ genes. This shows that when we are considering a reallocation of patients from one robust cluster to another within the complete Erasmus data ($n = 344$) we essentially have several thousand genes, that in reality do not discriminate between the three clusters, potentially affecting the decision.

10.3 Classification results

As we mentioned in the previous chapter there is a total of 62 patients with three known and well characterized subtypes of AML [sample sizes in brackets], inv(16) [$n_1 = 28$], $t(15;17)$ [$n_2 = 10$] and $t(8;21)$ [$n_3 = 24$]. We evaluated the performance of the classification procedure described in section 8.5 on these 62 patients using both the methylation and expression data, first separately and then jointly. We randomly split the 62 patients into a training set and a test set, ran the EM algorithm on the training set and then classified the test cases. We repeated this process $1,000$ times each time randomizing the patients into the training and test set. For each iteration we made sure that the same percentage of patients from each group was present in the training set. This

Table 10.4: The below Table summarizes the percentage of successfully classified test cases for the randomized classification experiment described in section 10.3.

% in training set	lik(meth)	lik(expr)	lik(meth&expr)
90%	97.2%	99.2%	100%
80%	96.8%	99.2%	100%
70%	96.4%	99.1%	99.9%

was done for percentage values (of patients in the training set) ranging from $70 - 90\%$. The classification success rates from this experiment are summarized in Table 10.4. We can see that the classification success rates based on the expression data are very good, around 99% for all percentage values, and better than the success rates based on the methylation data. As in the clustering case, the joint classification based on both data types simultaneously does better than each classification separately, with almost perfect success rate for each of the percentage values $70, 80$ and 90%.

10.4 Identifying discriminating genes

After we have found a partition of the patients set (or if we know the true partition) we wish to identify the genes that methylate differently across clusters. We run the likelihood algorithm on all the genes and in the process obtain estimates of the posterior expectations $(E[\gamma_j|\mathbf{y}, \boldsymbol{\theta}])_j$ and $(E[w_{cj}|\gamma_j = 1, \mathbf{y}, \boldsymbol{\theta}])_{c,j}$. It is clear that genes j with $\gamma_j = 0$ or $w_{cj} = w_{c'j}$ for all c, c' will not be declared significant. Hence we declare a gene j significant if $\gamma_j = 1$ and there exists a pair of distinct classes c, c' such that $w_{cj} \neq w_{c'j}$. For each gene j we can calculate the estimated posterior probability of the

event $S_j = [\gamma_j = 1] \cap [w_{cj} \neq w_{c'j}, \text{ for some } c \neq c']$:

$$\begin{aligned} p_j &= P\{S_j|\mathbf{y}\} \\ &= P\{[\gamma_j = 1] \cap [w_{cj} \neq w_{c'j}, \text{ for some } c \neq c']|\mathbf{y}\} \\ &= (1 - P\{w_{cj} = w_{c'j}, \text{ all } c, c'|\gamma_j = 1, \mathbf{y}\}) \cdot P\{\gamma_j = 1|\mathbf{y}\} \\ &= (1 - P\{w_{cj} = 1, \text{ all } c|\gamma_j = 1, \mathbf{y}\} - P\{w_{cj} = 0, \text{ all } c|\gamma_j = 1, \mathbf{y}\}) \cdot P\{\gamma_j = 1|\mathbf{y}\} \\ &= \Big(1 - \prod_{c=1}^{K} E[w_{cj}|\gamma_j = 1, \mathbf{y}] - \prod_{c=1}^{K}(1 - E[w_{cj}|\gamma_j = 1, \mathbf{y}])\Big) \cdot E[\gamma_j|\mathbf{y}]. \end{aligned}$$

We can now determine a significance threshold and declare a gene significant if p_j exceeds that threshold. Let us now rank the above probabilities and let $p_{(j)}$ denote the jth largest probability, $j = 1, \ldots, G$. Thus $p_{(1)}$ represents the largest significance probability among the G genes and $p_{(G)}$ the smallest. In Figure 10.7 we see a plot of the ordered probabilities $p_{(j)}$ against the ranks j (solid black curve) for the methylation data and the partition M_{full}+CEM (see section 10.2.5). Note that if we declare all genes j that fulfill $p_j > 1 - \delta$, for some $\delta > 0$, we are making a probabilistic statement about each of the significant genes separately. More specifically, we say for each of the declared significant genes that the estimated probability of it being a discriminating gene is greater than $1 - \delta$. However, using the above approach we cannot say that all of the declared significant genes are discriminating with estimated probability greater than $1 - \delta$. To make a global statement about the significant genes we need an alternative approach that uses the fact that the events $S_1, \ldots, S_G$ are all independent. We let $S_{(j)}$ denote the significance event corresponding to the gene with the jth largest probability $p_{(j)}$. For a given $\delta > 0$ we find the smallest G^* such that

$$P_{G^*} = P\Big(\bigcap_{j=1}^{G^*} S_{(j)}|\mathbf{y}\Big) = \prod_{j=1}^{G^*} p_{(j)} > 1 - \delta$$

This method declares the G^* highest ranked genes significant. The estimated probability that all the declared genes are discriminating is controlled to be greater than $1 - \delta$. In Figure 10.7 we see a plot of the cumulative products, P_{G^*}, against G^* (dashed red

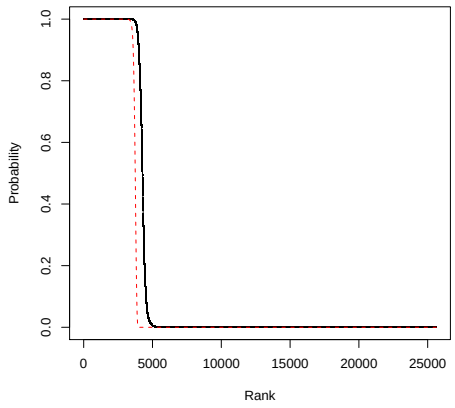

Figure 10.7: In the above plot we see a plot of the ordered probabilities, $p_{(G^*)}$, (black solid curve) and the cumulative products, $P_{G^*} = \prod_{j=1}^{G^*} p_{(j)}$, (red dashed curve) plotted against the ranks G^*. This is based on the methylation data and the partition M_{full}+CEM.

curve) for the methylation data and the partition M_{full}+CEM. For this partition and with $\delta = 0.001$, we declare $3,823$ genes significant using the criteria $p_j > 0.999$. If we however select the smallest G^* such that $P_{G^*} > 0.999$ we declare significant slightly less or $3,530$ genes. Upon arriving at a clustering result, Figueroa et al. (2010) performed an ANOVA on each of the $G = 25,626$ genes separately, and adjusted the p-values according to Benjamini and Hochberg (1995). Genes that passed the significance threshold were declared discriminating for the different subtypes. The above method, through the posterior expectations of the EM algorithm, automatically arrives at the discriminating genes. One major advantage over the ANOVA method is that we not only obtain information about which genes methylate differently across subtypes, but we can also through the matrix $(E[w_{cj}|\gamma_j = 1, \mathbf{y}])_{c,j}$ predict the nature of those differences. Furthermore, the results from the ANOVA should be taken with care due to the microarray effect, whereas our model based method adjusts for this array effect through the individual specific parameters.

CHAPTER 11

DISCUSSION

In this part of the thesis we have developed a likelihood based clustering scheme for clustering AML patients based on microarray data. The hierarchical algorithm of chapter 7 guides the selection of a good initial partition and the two way CEM algorithm can be used to improve upon this initial guess. The two way CEM algorithm of chapter 8 selects discriminating variables, alongside the clustering task, and thus filters out noise in the data in a model based manner. After having arrived at a candidate partition, we can run the extended likelihood model of chapter 8 and obtain posterior expectations of our gene-importance and methylation indicators. These posterior expectations can be used to guide the classification of new AML patients and furthermore provide us with information about which genes methylate differently across clusters. The latter tells us which genes discriminate between subtypes and thus there is no need to perform ANOVA on each gene after coming up with a candidate partition. Moreover we learn exactly which genes methylate and which don't in all clusters. This could become valuable information for understanding the biological bases of different subtypes of AML. Through the multiplatform model we can run an integrative analysis across multiple data types. The joint analysis of the methylation and expression data has been shown to achieve higher discriminating power than both of the single platform analyses separately. Overall, the likelihood based model gives some very promising results in clustering and classifying patients with AML. A future direction for this project is to run a fully Bayesian analysis using the more realistic model of chapter 9 as the building block. One of the hindrance of running a two way clustering on both genes and patients simultaneously was the infeasible form of the posterior density $f(\mathbf{w}, \boldsymbol{\gamma}|\mathbf{y})$. This lead us to the assumption of fixed methylation indicators, $\mathbf{w}$, and the two way CEM algorithm. The classification EM approach is known to give inconsistent estimators of the means

and variances (see McLachlan and Peel (2000)) and thus there might be some room for improvement through a fully Bayesian analysis. Due to the simple form of the full conditionals $f(\mathbf{w}|\boldsymbol{\gamma}, \mathbf{y})$ and $f(\boldsymbol{\gamma}|\mathbf{w}, \mathbf{y})$ a Bayesian Gibbs sampler might be a feasible option.

APPENDIX A

TECHNICAL DETAILS FROM PART I

A.1 Justifying the model for the hazard

Now we will attempt to justify the model (2.9), in particular the last equation, using the notion of counting processes. We will roughly follow the notation in chapter 5 of Kalbfleisch and Prentice (2002). Let T_i denote the failure time for patient i with continuous hazard function, $h_i(t)$, and assume there is an underlying counting process $\tilde{N}_i = \{\tilde{N}_i(t), t \geq 0\}$, $\tilde{N}_i(0) = 0$, where $\tilde{N}_i(t) = \text{I}(T_i \leq t)$ indicates whether patient i is alive, $\tilde{N}_i(t) = 0$, or dead, $\tilde{N}_i(t) = 1$, at or before time t. Let C_i denote censoring time and define $R_i(t) = \text{I}(T_i \geq t, C_i \geq t)$ to be the at risk process for individual i. With this definition the ith process is at risk of an observed event at time t if and only if $R_i(t) = 1$. We do not observe all the events in the underlying counting process $\tilde{N}_i$ if the individual is right censored. This leads us to defining the observed counting process $N_i = \{N_i(t), t \geq 0\}$, with $N_i(t) = \int_0^t R_i(u) d\tilde{N}_i(u)$, which denotes whether or not an event has occurred, and been observed, in the interval $(0, t]$. The actual data we collect at the beginning of each interval $[t_{j-1}, t_j)$ is whether or not the patient died or was censored in the previous interval and the endogenous measurement $Y_i(t_{j-1})$. We define $dN_i\left([t_{j-1}, t_j)\right) := N_i(t_j^-) - N_i(t_{j-1}^-)$ to be the number of observed events, 0 or 1, that occur in the jth interval $[t_{j-1}, t_j)$, $j = 1, \ldots, m$. This quantity approaches $dN_i(t) = N_i(t^- + dt) - N_i(t^-)$, the number of observed events in $[t, t + dt)$, as the partitioning becomes finer and finer with one of the partition points, $t_{j-1} = t$. We now define the patients' history up to the beginning of each interval, $[t_{j-1}, t_j)$:

$$\mathcal{F}_{t_{j-1}} = \sigma\{dN_i([t_{k-1}, t_k)), Y_i(t_{k-1}), R_i(t_k), X_i^F, i = 1, \ldots, n, k = 1, \ldots, j-1\}$$

which basically tells us whether the patient died in any of the previous intervals as well as the at risk indicator and the endogenous covariates at the beginning of each previous interval. Note that we also include the at risk indicator at the time of the visit t_{j-1}. We can think of this as the knowledge of whether the patient has actually physically arrived for the follow up visit. One could ask why $Y_i(t_{j-1})$ is not included, but we can assume that we have not taken the measurement $Y_i(t_{j-1})$ yet, and so it is not included in the patient's history. We define $X_i^F = \{X_i(u), 0 \leq u \leq \tau\}$ to be the full deterministic exogenous covariate history of patient i and note that

$$\mathcal{F}_0 = \mathcal{F}_{t_0} = \sigma\{R_i(0), X_i^F, i = 1, \ldots, n\} = \sigma\{X_i^F, i = 1, \ldots, n\}$$

as we can assume that everyone is at risk at the beginning of the study. With this filtration in mind we construct the likelihood. We first look at the data contribution to the likelihood in the interval $[0, t_1)$ conditioning on $\mathcal{F}_0$. Then conditioned on $\mathcal{F}_{t_1}$ we look at the data contribution to the likelihood in the interval $[t_1, t_2)$ and so on. Before we proceed however we need to make the two following assumption

$$\mathrm{P}\big[dN_i([t_{j-1}, t_j)) = 1 | Y_i(t_{j-1}), \mathcal{F}_{t_{j-1}}\big] = R_i(t_{j-1}) h_i([t_{j-1}, t_j)) \tag{A.1}$$

where

$$h_i([t_{j-1}, t_j)) := \mathrm{P}\big[T_i \in [t_{j-1}, t_j) | T_i \geq t_{j-1}, Y_i(t_{j-1}), \mathcal{F}_{t_{j-1}}\big] \tag{A.2}$$

is the discretized hazard of the jth interval. Note that $h_i([t_{j-1}, t_j))$ corresponds to $h_{i,j}$ in (2.8). We furthermore assume that $T_i | \mathcal{F}_{t_{j-1}}$ act independently over $[t_{j-1}, t_j)$. These two assumptions correspond to assumptions $1.$ and $2.$ in chapter 6.2 of Kalbfleisch and Prentice (2002). Now conditioning on $\mathcal{F}_{t_{j-1}}$ the data contribution to the interval $[t_{j-1}, t_j)$ is the triple $(dN_i([t_{j-1}, t_j)), R_i(t_j), Y_i(t_{j-1}))$. Note that there is an explicit assumption that for each interval we know whether or not a patient died or was censored within $[t_{j-1}, t_j)$. Hence knowing these three components gives us all the information about what happened within the interval. But this means that the likelihood contribution of the

interval $[t_{j-1}, t_j)$ can be factored into

$$\prod_{i=1}^{n} \mathrm{P}\big[R_i(t_j)|dN_i([t_{j-1}, t_j)), Y_i(t_{j-1}), \mathcal{F}_{t_{j-1}}\big] \tag{A.3}$$

$$\times \prod_{i=1}^{n} \mathrm{P}\big[dN_i([t_{j-1}, t_j))|Y_i(t_{j-1}), \mathcal{F}_{t_{j-1}}\big]\mathrm{P}\big[Y_i(t_{j-1})|\mathcal{F}_{t_{j-1}}\big] \tag{A.4}$$

For the first part note that by conditioning on $\mathcal{F}_{t_{j-1}}$ and $dN_i([t_{j-1}, t_j))$, we are provided with the value of $R_i(t_{j-1})$. If $dN_i([t_{j-1}, t_j)) = 1$ or $R_i(t_{j-1}) = 0$ we know that $R(t_j) = 0$ by definition. If $R_i(t_{j-1})) = 1$ and $dN_i([t_{j-1}, t_j)) = 0$ we know that $T_i \geq t_{j-1}$ and that there was no death in interval $[t_{j-1}, t_j)$, which means that $T_i \geq t_j$ as well. By this it is clear that given $(dN_i([t_{j-1}, t_j)), R_i(t_{j-1})) = (0, 1)$, the value of the random variable $R(t_j)$ is determined solely by whether the patient was censored in the interval $R(t_j) = 0$ or not $R(t_j) = 1$. Thus it is clear that the second part of the likelihood only involves the probabilistic structure of the censoring time and if we assume that the censoring is non-informative in the sense that $(A.3)$ is not a function of the parameters of interest we can base our inference on the following likelihood function

$$L = \prod_{j=1}^{m}\prod_{i=1}^{n} \mathrm{P}\big[dN_i([t_{j-1}, t_j))|Y_i(t_{j-1}), \mathcal{F}_{t_{j-1}}\big]\mathrm{P}\big[Y_i(t_{j-1})|\mathcal{F}_{t_{j-1}}\big]$$

where

$$\begin{aligned} & \mathrm{P}\big[dN_i([t_{j-1}, t_j))|Y_i(t_{j-1}), \mathcal{F}_{t_{j-1}}\big] \\ = \;& h_i([t_{j-1}, t_j))^{dN_i([t_{j-1}, t_j))}(1 - h_i([t_{j-1}, t_j)))^{R_i(t_{j-1}) - dN_i([t_{j-1}, t_j))} \end{aligned}$$

A.2 Independence of errors in a directed acyclic graph

In chapter 2 we noted that if the right hand sides of (2.2) are interpreted as conditional means plus error terms the directed acyclic graph structure imposes independence on

the errors. To get an intuitive understanding of the issue consider the simple model:

$$\begin{aligned} Y_1 &= \beta_1 x + \varepsilon_1 \\ Y_2 &= \beta_2 Y_1 + \varepsilon_2 \end{aligned}$$

where x is assumed nonrandom. We assume that the covariance matrix of ε_1 and ε_2 is

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

From the implied joint bivariate distribution of (Y_1, Y_2) it follows that

$$\mathrm{E}\left[Y_2|Y_1\right] = \beta_2 Y_1 + \frac{\sigma_{12}}{\sigma_1^2}(Y_1 - \beta_1 x) = (\beta_2 + \frac{\sigma_{12}}{\sigma_1^2})Y_1 - (\frac{\sigma_{12}}{\sigma_1^2}\beta_1)x$$

How do we interpret β_2 in the simple model above? If we were to think of $\beta_2 Y_1$ in the above model as the conditional mean as we do in the classical simple linear model it follows that

$$\mathrm{E}\left[Y_1 \varepsilon_2\right] = \mathrm{E}\left[Y_1(Y_2 - \mathrm{E}[Y_2|Y_1])\right] = 0$$

so Y_1 and ε_2 are independent, implying that ε_1 and ε_2 are independent, i.e. $\sigma_{12} = 0$.

Let us now consider the Dynamic path analysis model:

$$\begin{aligned} Y_{i1}(t) &= \mathbf{X}_i^{(1)}(t)\boldsymbol{\delta}_1(t) + \varepsilon_{i1}(t) \\ Y_{i2}(t) &= \mathbf{Y}_i^{(2)}(t)\boldsymbol{\gamma}_2(t) + \mathbf{X}_i^{(2)}(t)\boldsymbol{\delta}_2(t) + \varepsilon_{i2}(t) \\ &\vdots \\ Y_{iG}(t) &= \mathbf{Y}_i^{(G)}(t)\boldsymbol{\gamma}_G(t) + \mathbf{X}_i^{(G)}(t)\boldsymbol{\delta}_G(t) + \varepsilon_{iG}(t) \\ Y_{i,G+1}(t) &= \mathbf{Y}_i^{(G+1)}(t)\boldsymbol{\gamma}_{G+1}(t) + \mathbf{X}_i^{(G+1)}(t)\boldsymbol{\delta}_{G+1}(t) + \varepsilon_{i,G+1}(t) \end{aligned}$$

If we think of the linear predictors in each equation as conditional means it implies that

the ε's are independent. Write

$$\begin{aligned} Y_{i1}(t) &= \mathrm{E}\left[Y_{i1}(t)|\mathbf{X}\right] + \varepsilon_{i1}(t) \\ Y_{i2}(t) &= \mathrm{E}\left[Y_{i2}(t)|Y_{i1}(t), \mathbf{X}\right] + \varepsilon_{i2}(t) \\ &\vdots \\ Y_{i,G+1}(t) &= \mathrm{E}\left[Y_{i,G+1}(t)|Y_{i1}(t), \ldots, Y_{i,G}, \mathbf{X}\right] + \varepsilon_{i,G+1}(t) \end{aligned}$$

It follows for $j < k$ that

$$\begin{aligned} \mathrm{E}\left\{\varepsilon_j \varepsilon_k\right\} &= \mathrm{E}\big[(Y_j - \mathrm{E}\left[Y_j|Y_1, \ldots, Y_{j-1}, \mathbf{X}\right])(Y_k - \mathrm{E}\left[Y_k|Y_1, \ldots, Y_{k-1}, \mathbf{X}\right])\big] \\ &= \mathrm{E}\left[Y_j Y_k\right] - \mathrm{E}\big[Y_j \mathrm{E}\left[Y_k|Y_1, \ldots, Y_{k-1}, \mathbf{X}\right]\big] && \text{(A.5)} \\ &\quad -\mathrm{E}\big[Y_k \mathrm{E}\left[Y_j|Y_1, \ldots, Y_{j-1}, \mathbf{X}\right]\big] && \text{(A.6)} \\ &\quad +\mathrm{E}\big[\mathrm{E}\left[Y_j|Y_1, \ldots, Y_{j-1}, \mathbf{X}\right] \mathrm{E}\left[Y_k|Y_1, \ldots, Y_{k-1}, \mathbf{X}\right]\big] && \text{(A.7)} \\ &= 0 \end{aligned}$$

The terms in $(A.5)$ cancel since Y_j is measurable with respect to the σ-algebra generated by $Y_1, \ldots, Y_{k-1}, \mathbf{X}$. The term in $(A.7)$ cancels with the term in $(A.6)$ since $\mathrm{E}\left[Y_j|Y_1, \ldots, Y_{j-1}, \mathbf{X}\right]$ is measurable with respect to the σ-algebra generated by $Y_1, \ldots, Y_{k-1}, \mathbf{X}$.

A.3 Proof of Lemmas $(3.2.1)$ and $(3.2.2)$

A.3.1 Proof of Lemma $(3.2.1)$

Let us derive the likelihood of $\mathbf{Y}$. The error structure in (3.13) can be written in matrix form in the following way

$$\begin{aligned} \boldsymbol{\varepsilon}_i(t_{i1}) &= \mathbf{e}_i(t_{i1}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_1) \\ \boldsymbol{\varepsilon}_i(t_{ij}) &= \boldsymbol{\Phi}\boldsymbol{\varepsilon}_i(t_{i,j-1}) + \mathbf{e}_i(t_{ij}), \quad j \geq 2 \end{aligned} \tag{A.8}$$

where $\boldsymbol{\Phi} = \mathrm{diag}_{1\leq k\leq G+1}(\phi_k)$ and $\mathbf{e}_i(t_{ij}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_j)$, $j \geq 2$. With this we can plug into $(5!)$ and get

$$\mathbf{Y}_i(t_{i1}) = \mathbf{W}_i(t_{i1})\boldsymbol{\beta} + \mathbf{Z}_i(t_{i1})\mathbf{u} + \mathbf{e}(t_{i1})$$

and for $j \geq 2$

$$\begin{aligned} \mathbf{Y}_i(t_{ij}) &= \mathbf{W}_i(t_{ij})\boldsymbol{\beta} + \mathbf{Z}_i(t_{ij})\mathbf{u} \\ &\quad + \boldsymbol{\Phi}\left(\mathbf{Y}_i(t_{i,j-1}) - \mathbf{W}_i(t_{i,j-1})\boldsymbol{\beta} - \mathbf{Z}_i(t_{i,j-1})\mathbf{u}\right) + \mathbf{e}(t_{ij}) \end{aligned}$$

Recall that we have $\mathbf{W}_i(t)\boldsymbol{\beta} + \mathbf{Z}_i(t)\mathbf{u} = \boldsymbol{\Gamma}(t)\mathbf{Y}_i(t) + \boldsymbol{\Delta}(t)\mathbf{X}_i(t)$ but the latter form we'll use in the derivation of the density of $\mathbf{Y}$. The above equations become

$$\mathbf{Y}_i(t_{i1}) = (\mathbf{I} - \boldsymbol{\Gamma}(t_{i1}))^{-1}\boldsymbol{\Delta}(t_{i1})\mathbf{X}_i(t_{ij}) + (\mathbf{I} - \boldsymbol{\Gamma}(t_{i1}))^{-1}\mathbf{e}(t_{i1})$$

and for all $j \geq 2$

$$\begin{aligned} \mathbf{Y}_i(t_{ij}) &= (\mathbf{I} - \boldsymbol{\Gamma}(t_{ij}))^{-1}\Bigg(\boldsymbol{\Delta}(t_{ij})\mathbf{X}_i(t_{ij}) + \boldsymbol{\Phi}\Big(\mathbf{Y}_i(t_{i,j-1}) \\ &\quad -\boldsymbol{\Gamma}(t_{i,j-1})\mathbf{Y}_i(t_{i,j-1}) - \boldsymbol{\Delta}(t_{i,j-1})\mathbf{X}_i(t_{i,j-1})\Big)\Bigg) + (\mathbf{I} - \boldsymbol{\Gamma}(t_{ij}))^{-1}\mathbf{e}(t_{ij}) \end{aligned}$$

It is easy to verify that $\mathbf{Y}_i$ are independent for all $i = 1, \ldots, n$ and so using the above the likelihood factors into

$$\begin{aligned} \big[\mathbf{y}|\boldsymbol{\Theta}\big] &= \prod_{i=1}^{n} \left(\big[\mathbf{y}_i(t_{i1})|\boldsymbol{\Theta}\big] \prod_{j=2}^{m_i} \big[\mathbf{y}_i(t_{ij})|\mathbf{y}_i(t_{i,j-1}), \boldsymbol{\Theta}\big] \right) \end{aligned} \tag{A.9}$$

where

$$\begin{aligned} & \mathbf{y}_i(t_{i1})|\boldsymbol{\Theta} \\ \sim\ & \mathrm{N}\left((\mathbf{I} - \boldsymbol{\Gamma}(t_{i1}))^{-1}\boldsymbol{\Delta}(t_{i1})\mathbf{X}_i(t_{ij}), (\mathbf{I} - \boldsymbol{\Gamma}(t_{i1}))^{-1}\boldsymbol{\Sigma}_1(\mathbf{I} - \boldsymbol{\Gamma}(t_{i1}))^{-T}\right) \end{aligned} \tag{A.10}$$

with density

$$\begin{aligned} \big[\mathbf{y}_i(t_{i1})|\boldsymbol{\Theta}\big] &= (2\pi)^{-(G+1)/2}|\boldsymbol{\Sigma}_1|^{-1/2} \exp\left(\left((\mathbf{I} - \boldsymbol{\Gamma}(t_{i1}))\mathbf{Y}_i(t_{i1}) - \boldsymbol{\Delta}(t_{i1})\mathbf{X}_i(t_{i1}) \right)' \right. \\ &\quad \left. \times \boldsymbol{\Sigma}_1^{-1} \left((\mathbf{I} - \boldsymbol{\Gamma}(t_{i1}))\mathbf{Y}_i(t_{i1}) - \boldsymbol{\Delta}(t_{i1})\mathbf{X}_i(t_{i1}) \right) \right) \\ &= (2\pi)^{-(G+1)/2}|\boldsymbol{\Sigma}_1|^{-1/2} e^{-\frac{1}{2}(\mathbf{y}_i^*(t_{i1}) - \boldsymbol{\mu}_i^*(t_{i1}))'\boldsymbol{\Sigma}_1^{-1}(\mathbf{y}_i^*(t_{i1}) - \boldsymbol{\mu}_i^*(t_{i1}))} \end{aligned} \tag{A.11}$$

and for $j \geq 2$

$$\begin{aligned} & \mathbf{y}_i(t_{ij})|\mathbf{y}_i(t_{i,j-1}), \boldsymbol{\Theta} \\ \sim\ & \mathrm{N}\Bigg((\mathbf{I} - \boldsymbol{\Gamma}(t_{ij}))^{-1} \Big(\boldsymbol{\Delta}(t_{ij})\mathbf{X}_i(t_{ij}) + \boldsymbol{\Phi}\{\mathbf{Y}_i(t_{i,j-1}) - \boldsymbol{\Gamma}(t_{i,j-1})\mathbf{Y}_i(t_{i,j-1}) \\ & - \boldsymbol{\Delta}(t_{i,j-1})\mathbf{X}_i(t_{i,j-1})\} \Big), (\mathbf{I} - \boldsymbol{\Gamma}(t_{ij}))^{-1}\boldsymbol{\Sigma}_j(\mathbf{I} - \boldsymbol{\Gamma}(t_{ij}))^{-T} \Bigg) \end{aligned} \tag{A.12}$$

with the same density as that of $(\mathbf{I}-\mathbf{\Gamma}(t_{ij}))\mathbf{y}_i(t_{ij})|\mathbf{y}_i(t_{i,j-1}),\mathbf{\Theta}$, because of the directed acyclic graph structure, or

$$
\begin{aligned}
& \big[\mathbf{y}_i(t_{ij})|\mathbf{y}_i(t_{i,j-1}),\mathbf{\Theta}\big] \\
= \;& (2\pi)^{-(G+1)/2}|\mathbf{\Sigma}_j|^{-1/2}\exp\Bigg(\Bigg[(\mathbf{I}-\mathbf{\Gamma}(t_{ij}))\mathbf{Y}_i(t_{ij})-\mathbf{\Delta}(t_{ij})\mathbf{X}_i(t_{ij}) \\
& -\mathbf{\Phi}\Big((\mathbf{I}-\mathbf{\Gamma}(t_{i,j-1}))\mathbf{Y}_i(t_{i,j-1})-\mathbf{\Delta}(t_{i,j-1})\mathbf{X}_i(t_{i,j-1})\Big)\Bigg]' \\
& \times\mathbf{\Sigma}_j^{-1}\Bigg[(\mathbf{I}-\mathbf{\Gamma}(t_{ij}))\mathbf{Y}_i(t_{ij})-\mathbf{\Delta}(t_{ij})\mathbf{X}_i(t_{ij}) \\
& -\mathbf{\Phi}\Big((\mathbf{I}-\mathbf{\Gamma}(t_{i,j-1}))\mathbf{Y}_i(t_{i,j-1})-\mathbf{\Delta}(t_{i,j-1})\mathbf{X}_i(t_{i,j-1})\Big)\Bigg]\Bigg) \\
= \;& (2\pi)^{-(G+1)/2}|\mathbf{\Sigma}_j|^{-1/2}\exp\Bigg(\Bigg[\Big(\mathbf{Y}_i(t_{ij})-\mathbf{\Phi}\mathbf{Y}_i(t_{i,j-1})\Big) \\
& -\Big(\mathbf{W}_i(t_{ij})\boldsymbol{\beta}+\mathbf{Z}_i(t_{ij})\mathbf{u}-\mathbf{\Phi}\{\mathbf{W}_i(t_{i,j-1})\boldsymbol{\beta}+\mathbf{Z}_i(t_{i,j-1})\mathbf{u}\}\Big)\Bigg]' \\
& \times\mathbf{\Sigma}_j^{-1}\Bigg[\Big(\mathbf{Y}_i(t_{ij})-\mathbf{\Phi}\mathbf{Y}_i(t_{i,j-1})\Big) \\
& -\Big(\mathbf{W}_i(t_{ij})\boldsymbol{\beta}+\mathbf{Z}_i(t_{ij})\mathbf{u}-\mathbf{\Phi}\{\mathbf{W}_i(t_{i,j-1})\boldsymbol{\beta}+\mathbf{Z}_i(t_{i,j-1})\mathbf{u}\}\Big)\Bigg]\Bigg) \\
= \;& (2\pi)^{-(G+1)/2}|\mathbf{\Sigma}_j|^{-1/2}e^{-\frac{1}{2}\left(\mathbf{y}_i^*(t_{ij})-\boldsymbol{\mu}_i^*(t_{ij})\right)'\mathbf{\Sigma}_j^{-1}\left(\mathbf{y}_i^*(t_{ij})-\boldsymbol{\mu}_i^*(t_{ij})\right)} \qquad \text{(A.13)}
\end{aligned}
$$

Using $(A.9)$, $(A.11)$ and $(A.13)$ we see that the likelihood is equal to

$$
(2\pi)^{-(G+1)m/2}\Big(\prod_{i=1}^{n}\prod_{j=1}^{m_i}|\mathbf{\Sigma}_j|^{-1/2}\Big)e^{-\frac{1}{2}\sum_{i,j}(\mathbf{y}_i^*(t_{ij})-\boldsymbol{\mu}_i^*(t_{ij}))'\mathbf{\Sigma}_j^{-1}(\mathbf{y}_i^*(t_{ij})-\boldsymbol{\mu}_i^*(t_{ij}))} \qquad \text{(A.14)}
$$

Using the definition of $\mathbf{\Omega}=\text{blockdiag}_{1\le i\le n}(\text{blockdiag}_{1\le j\le m_i}(\mathbf{\Sigma}_j))$ the likelihood can be written more compactly

$$
\big[\mathbf{y}^*|\mathbf{\Theta}\big] \;=\; (2\pi)^{-(G+1)m/2}|\mathbf{\Omega}|^{-1/2}e^{-\frac{1}{2}(\mathbf{y}^*-\mathbf{W}^*\boldsymbol{\beta}-\mathbf{Z}^*\mathbf{u})'\mathbf{\Omega}^{-1}(\mathbf{y}^*-\mathbf{W}^*\boldsymbol{\beta}-\mathbf{Z}^*\mathbf{u})}
$$

which is the desired result.

A.3.2 Proof of Lemma $(3.2.2)$

It is easy to see that the density factors into

$$\begin{aligned}
\left[\mathbf{s}|\mathbf{y}_{G+1},\mathbf{y}_{obs},\mathbf{\Theta}\right] &= \left[(s_i(t_{ij}))_{i,j}|\mathbf{y}_{G+1},\mathbf{\Theta}\right] \\
&= \prod_{i=1}^{n}\left[\mathrm{I}(Y_{i,G+1}(t_{i1})>0),\ldots,\mathrm{I}(Y_{i,G+1}(t_{im_i})>0)|\mathbf{y}_{i,G+1},\mathbf{\Theta}\right] \\
&= \prod_{i=1}^{n}\Big(\left[\mathrm{I}(Y_{i,G+1}(t_{i1})>0)|\mathbf{y}_{i,G+1},\mathbf{\Theta}\right] \\
&\quad \times\prod_{j=2}^{m_i}\left[\mathrm{I}(Y_{i,G+1}(t_{ij})>0)|\mathrm{I}(Y_{i,G+1}(t_{i,j-1})>0),\mathbf{y}_{i,G+1},\mathbf{\Theta}\right]\Big) \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m_i}\left[\mathrm{I}(Y_{i,G+1}(t_{ij})>0)|y_{i,G+1}(t_{ij})\right] \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m_i}\mathrm{I}(y_{i,G+1}(t_{ij})\in A_{ij}(s_i(t_{ij}))
\end{aligned}$$

where the last equality was already established in the independent case.

APPENDIX B

TECHNICAL DETAILS FROM PART II

B.1 EM algorithm of chapter 7

Recall that in the M-step of the EM algorithm of section 7.2 we needed to maximize the Q_c-function,

$$\begin{aligned} Q_c(\boldsymbol{\theta}_c|\boldsymbol{\theta}_c^{(t)}) &= \sum_{j\in J_d}\Big(\tau_{cj}^{(t)}\log\pi_{1c} + \big(1-\tau_{cj}^{(t)}\big)\log\pi_{0c}\Big) \\ &\quad + \sum_{j\in J_d}\left(\tau_{cj}^{(t)}\sum_{i\in c}\log\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2) + \big(1-\tau_{cj}^{(t)}\big)\sum_{i\in c}\log\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2)\right), \end{aligned}$$

with respect to $((\pi_{1c}),(\mu_{1i},\sigma_{1i}^2,\mu_{2i},\sigma_{2i}^2)_{i\in c})$, where

$$\log\phi(y_{ij}|\mu_{ki},\sigma_{ki}^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_{ki}^2) - \frac{1}{2\sigma_{ki}^2}(y_{ij}-\mu_{ki})^2, \quad k=1,2.$$

Differentiating with respect to π_{1c} and setting to zero we get (recall $G_\delta = \#J_d$)

$$\sum_{j=1}^{G_\delta}\left(\frac{\tau_{cj}^{(t)}}{\pi_{1c}} - \frac{1-\tau_{cj}^{(t)}}{\pi_{0c}}\right) = 0,$$

which leads to

$$\frac{\sum_{j=1}^{G_\delta}\tau_{cj}^{(t)}}{\pi_{1c}} = \frac{G_\delta - \sum_{j=1}^{G_\delta}\tau_{cj}^{(t)}}{1-\pi_{1c}} = G_\delta,$$

so

$$\pi_{1c}^{(t+1)} = \frac{1}{G_\delta}\sum_{j=1}^{G_\delta}\tau_{cj}^{(t)}.$$

Differentiating with respect to μ_{1i}, for $i\in c$, and setting to zero we get

$$\sum_{j=1}^{G_\delta}\frac{\tau_{cj}^{(t)}}{\sigma_{1i}^2}\big(y_{ij}-\mu_{1i}\big) = 0,$$

which leads to

$$\mu_{1i}^{(t+1)} = \frac{\sum_{j=1}^{G_\delta}\tau_{cj}^{(t)}y_{ij}}{\sum_{j=1}^{G_\delta}\tau_{cj}^{(t)}}.$$

Similarly if we differentiate with respect to μ_{2i}, for $i \in c$, and set to zero we get

$$\mu_{2i}^{(t+1)} = \frac{\sum_{j=1}^{G_\delta}(1-\tau_{cj}^{(t)})y_{ij}}{\sum_{j=1}^{G_\delta}(1-\tau_{cj}^{(t)})}.$$

Now differentiating with respect to σ_{1i}^2, for $i \in c$, and setting to zero we get

$$-\sum_{j=1}^{G_\delta}\frac{\tau_{cj}^{(t)}}{2\sigma_{1i}^2} + \sum_{j=1}^{G_\delta}\frac{\tau_{cj}^{(t)}(y_{ij}-\mu_{1i}^{(t+1)})^2}{2(\sigma_{1i}^2)^2} = 0,$$

which leads to

$$\sigma_{1i}^{2(t+1)} = \frac{\sum_{j=1}^{G_\delta}\tau_{cj}^{(t)}\left(y_{ij}-\mu_{1i}^{(t+1)}\right)^2}{\sum_{j=1}^{G_\delta}\tau_{cj}^{(t)}}.$$

Finally, if we differentiate with respect to σ_{2i}^2, for $i \in c$, and set to zero we get

$$\sigma_{2i}^{2(t+1)} = \frac{\sum_{j=1}^{G_\delta}\left(1-\tau_{cj}^{(t)}\right)\left(y_{ij}-\mu_{2i}^{(t+1)}\right)^2}{\sum_{j=1}^{G_\delta}\left(1-\tau_{cj}^{(t)}\right)}$$

B.2 EM algorithm of chapter 8

B.2.1 E-step

We now derive the posterior densities $f(\mathbf{w}|\boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\theta}^{(t)})$ and $f(\boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\theta}^{(t)})$ from section 8.2. Let us first verify the density $f(\mathbf{y}, \boldsymbol{\gamma})$ by integrating out $\mathbf{w}$ from

$$\begin{aligned} & f(\mathbf{y}, \mathbf{w}, \boldsymbol{\gamma}) \\ = & \prod_{j=1}^{G}\left(p\prod_{c\in\mathcal{C}}\left\{\pi_{1c}\prod_{i\in c}\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2)\right\}^{w_{cj}}\left\{\pi_{0c}\prod_{i\in c}\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2)\right\}^{1-w_{cj}}\right)^{\gamma_j} \qquad \text{(B.1)} \\ \times & \left((1-p)I(\mathbf{w}_j\in A)\left\{\pi_1\prod_{i=1}^{n}\phi(y_{ij}|\alpha_{1i},\varsigma_{1i}^2)\right\}^{w_{1j}}\left\{\pi_0\prod_{i=1}^{n}\phi(y_{ij}|\alpha_{1i},\varsigma_{1i}^2)\right\}^{1-w_{1j}}\right)^{1-\gamma_j}, \end{aligned}$$

where we recall $A = \{\mathbf{w}_j | w_{cj} = w_{c'j}, \text{ all } c, c' \in \mathcal{C}\}$. Fix j and sum over all possible values of $(w_{cj})_{c\in\mathcal{C}} \in \{0,1\}^K$, where $K = \#\mathcal{C}$. Let's for ease of notation in the derivation assume that the names of the clusters are $c = \{1\}, \ldots, \{K\}$, with methylation indicators $w_{1j}, \ldots, w_{Kj}$ respectively. Start by summing over $w_{1j} = 0, 1$:

$$
\begin{aligned}
&\sum_{\mathbf{w}_{1j}=0}^{1} \Bigg\{ \Bigg(p \prod_{c\in\mathcal{C}} \Big\{ \pi_{1c} \prod_{i\in c} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \Big\}^{w_{cj}} \Big\{ \pi_{0c} \prod_{i\in c} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) \Big\}^{1-w_{cj}} \Bigg)^{\gamma_j} \\
&\times \Big((1-p) I(\mathbf{w}_j \in A) \big\{ \pi_1 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}, \varsigma_{1i}^2) \big\}^{w_{1j}} \big\{ \pi_0 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}, \varsigma_{1i}^2) \big\}^{1-w_{1j}} \Big)^{1-\gamma_j} \Bigg\} \\
= \quad & \Bigg(p \Big(\pi_{11} \prod_{i\in\{1\}} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \Big) \\
&\times \prod_{c=2}^{K} \Big\{ \pi_{1c} \prod_{i\in\{c\}} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \Big\}^{w_{cj}} \Big\{ \pi_{0c} \prod_{i\in\{c\}} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) \Big\}^{1-w_{cj}} \Bigg)^{\gamma_j} \\
&\times \Bigg((1-p) I(1 = w_{2j} = \cdots = w_{Kj}) \Big\{ \pi_1 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}, \varsigma_{1i}^2) \Big\} \Bigg)^{1-\gamma_j} \\
+ \quad & \Bigg(p \Big(\pi_{01} \prod_{i\in\{1\}} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) \Big) \\
&\times \prod_{c=2}^{K} \Big\{ \pi_{1c} \prod_{i\in\{c\}} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \Big\}^{w_{cj}} \Big\{ \pi_{0c} \prod_{i\in\{c\}} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) \Big\}^{1-w_{cj}} \Bigg)^{\gamma_j} \\
&\times \Bigg((1-p) I(0 = w_{2j} = \cdots = w_{Kj}) \Big\{ \pi_0 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{2i}, \varsigma_{2i}^2) \Big\} \Bigg)^{1-\gamma_j} \\
= \quad & \Bigg(p \Big(\pi_{11} \prod_{i\in\{1\}} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) + \pi_{01} \prod_{i\in\{1\}} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) \Big) \\
&\times \prod_{c=2}^{K} \Big\{ \pi_{1c} \prod_{i\in\{c\}} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \Big\}^{w_{cj}} \Big\{ \pi_{0c} \prod_{i\in\{c\}} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2) \Big\}^{1-w_{cj}} \Bigg)^{\gamma_j} \\
\times \quad & \Bigg((1-p) I(1 = w_{2j} = \cdots = w_{Kj}) \Big\{ \pi_1 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}, \varsigma_{1i}^2) \Big\} \\
&\quad + (1-p) I(0 = w_{2j} = \cdots = w_{Kj}) \Big\{ \pi_0 \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{2i}, \varsigma_{2i}^2) \Big\} \Bigg)^{1-\gamma_j}
\end{aligned}
$$

where the last equality is easily verified for both $\gamma_j = 1$ and $\gamma_j = 0$. We continue in a similar fashion, sum next over $w_{2j} = 0, 1$, then $w_{3j} = 0, 1$ until we finally sum over

$w_{Kj} = 0, 1$. It's easy to see that we arrive at

$$\begin{aligned} f(\mathbf{y}_j, \gamma_j) &= \left\{ p \prod_{c \in \mathcal{C}} \Big(\pi_{1c} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) + \pi_{0c} \prod_{i \in c} \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2) \Big) \right\}^{\gamma_j} \\ &\quad \times \left\{ (1-p) \Big(\pi_1 \prod_{i=1}^{n} \phi(y_{ij} | \alpha_{1i}, \varsigma_{1i}^2) + \pi_0 \prod_{i=1}^{n} \phi(y_{ij} | \alpha_{2i}, \varsigma_{2i}^2) \Big) \right\}^{1-\gamma_j}, \end{aligned} \tag{B.2}$$

and we have

$$f(\mathbf{y}, \boldsymbol{\gamma}) = \prod_{j=1}^{G} f(\mathbf{y}_j, \gamma_j).$$

If we now sum (B.2) over $\gamma_j = 0, 1$, for each j, we get

$$\begin{aligned} f(\mathbf{y}) &= \prod_{j=1}^{G} \left\{ p \prod_{c \in \mathcal{C}} \Big(\pi_{1c} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}, \sigma_{1i}^2) + \pi_{0c} \prod_{i \in c} \phi(y_{ij} | \mu_{2i}, \sigma_{2i}^2) \Big) \right. \\ &\quad \left. + (1-p) \Big(\pi_1 \prod_{i=1}^{n} \phi(y_{ij} | \alpha_{1i}, \varsigma_{1i}^2) + \pi_0 \prod_{i=1}^{n} \phi(y_{ij} | \alpha_{2i}, \varsigma_{2i}^2) \Big) \right\}. \end{aligned} \tag{B.3}$$

It is now easy to derive the posterior densities $f(\mathbf{w}_j | \boldsymbol{\gamma}_j, \mathbf{y}_j, \boldsymbol{\theta}^{(t)})$ and $f(\boldsymbol{\gamma}_j | \mathbf{y}_j, \boldsymbol{\theta}^{(t)})$:

Theorem B.2.1. *Conditioning on the gene importance indicator, γ_j, the posterior density of $\mathbf{w}_j$ is*

$$\begin{aligned} f(\mathbf{w}_j | \gamma_j, \mathbf{y}_j, \boldsymbol{\theta}^{(t)}) &= \Big(\prod_{c \in \mathcal{C}} \big(\tau_{cj}^{(t)}\big)^{w_{cj}} \big(1 - \tau_{cj}^{(t)}\big)^{1-w_{cj}} \Big)^{\gamma_j} \\ &\quad \times \Big(I_A(\mathbf{w}_j) \big(\nu_{1j}^{(t)}\big)^{w_{1j}} \big(1 - \nu_{1j}^{(t)}\big)^{1-w_{1j}} \Big)^{1-\gamma_j}, \end{aligned} \tag{B.4}$$

where we recall the definitions of $\tau_{cj}^{(t)}$ and $\nu_{1j}^{(t)}$ from (8.9)

$$\tau_{cj}^{(t)} = \frac{\pi_{1c}^{(t)} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}^{(t)}, \sigma_{1i}^{2(t)})}{\pi_{1c}^{(t)} \prod_{i \in c} \phi(y_{ij} | \mu_{1i}^{(t)}, \sigma_{1i}^{2(t)}) + \pi_{0c}^{(t)} \prod_{i \in c} \phi(y_{ij} | \mu_{2i}^{(t)}, \sigma_{2i}^{2(t)})}, \tag{B.5}$$

$$\nu_{1j}^{(t)} = \frac{\pi_{1}^{(t)} \prod_{i=1}^{n} \phi(y_{ij} | \alpha_{1i}^{(t)}, \varsigma_{1i}^{2(t)})}{\pi_{1}^{(t)} \prod_{i=1}^{n} \phi(y_{ij} | \alpha_{1i}^{(t)}, \varsigma_{1i}^{2(t)}) + \pi_{0}^{(t)} \prod_{i=1}^{n} \phi(y_{ij} | \alpha_{2i}^{(t)}, \varsigma_{2i}^{2(t)})}. \tag{B.6}$$

Proof. From (B.1) and (B.2) we get

$$
\begin{aligned}
& f(\mathbf{w}_j|\gamma_j, \mathbf{y}_j, \boldsymbol{\theta}^{(t)}) \\
= & \frac{f(\mathbf{y}_j, \mathbf{w}_j, \gamma_j|\boldsymbol{\theta}^{(t)})}{f(\mathbf{y}_j, \gamma_j|\boldsymbol{\theta}^{(t)})} \\
= & \frac{\left(\prod_{c\in\mathcal{C}} \left\{\pi_{1c}^{(t)} \prod_{i\in c} \phi(y_{ij}|\mu_{1i}^{(t)}, \sigma_{1i}^{2(t)})\right\}^{w_{cj}} \left\{\pi_{0c}^{(t)} \prod_{i\in c} \phi(y_{ij}|\mu_{2i}^{(t)}, \sigma_{2i}^{2(t)})\right\}^{1-w_{cj}}\right)^{\gamma_j}}{\left(\prod_{c\in\mathcal{C}} \left(\pi_{1c}^{(t)} \prod_{i\in c} \phi(y_{ij}|\mu_{1i}^{(t)}, \sigma_{1i}^{2(t)}) + \pi_{0c}^{(t)} \prod_{i\in c} \phi(y_{ij}|\mu_{2i}^{(t)}, \sigma_{2i}^{2(t)})\right)\right)^{\gamma_j}} \\
\times & \frac{\left(I_A(\mathbf{w}_j)\left\{\pi_1^{(t)} \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}^{(t)}, \varsigma_{1i}^{2(t)})\right\}^{w_{1j}} \left\{\pi_0^{(t)} \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}^{(t)}, \varsigma_{1i}^{2(t)})\right\}^{1-w_{1j}}\right)^{1-\gamma_j}}{\left(\left(\pi_1^{(t)} \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{1i}^{(t)}, \varsigma_{1i}^{2(t)}) + \pi_0^{(t)} \prod_{i=1}^{n} \phi(y_{ij}|\alpha_{2i}^{(t)}, \varsigma_{2i}^{2(t)})\right)\right)^{1-\gamma_j}} \\
= & \left(\prod_{c\in\mathcal{C}} \left(\tau_{cj}^{(t)}\right)^{w_{cj}} \left(1-\tau_{cj}^{(t)}\right)^{1-w_{cj}}\right)^{\gamma_j} \left(I_A(\mathbf{w}_j) \left(\nu_{1j}^{(t)}\right)^{w_{1j}} \left(1-\nu_{1j}^{(t)}\right)^{1-w_{1j}}\right)^{1-\gamma_j},
\end{aligned}
$$

□

Lemma B.2.2. *Conditioning on the gene importance indicator, γ_j, the posterior expectation of w_{cj}, with respect to the density $f(\mathbf{w}_j|\boldsymbol{\gamma}_j, \mathbf{y}_j, \boldsymbol{\theta}^{(t)})$, is*

$$
E[w_{cj}|\gamma_j, \mathbf{y}_j, \boldsymbol{\theta}^{(t)}] = \gamma_j \tau_{cj}^{(t)} + (1-\gamma_j)\nu_{1j}^{(t)}.
$$

Proof. Using the definitions in (B.5) and (B.6) and the posterior density in (B.4) we get

$$
\begin{aligned}
E[w_{cj}|\gamma_j = 1, \mathbf{y}_j, \boldsymbol{\theta}^{(t)}] &= \int \cdots \int w_{cj} \prod_{c'\in\mathcal{C}} \left(\tau_{c'j}^{(t)}\right)^{w_{c'j}} \left(1-\tau_{c'j}^{(t)}\right)^{1-w_{c'j}} d\mathbf{w}_j \\
&= \tau_{cj}^{(t)},
\end{aligned}
$$

and

$$
\begin{aligned}
E[w_{cj}|\gamma_j = 0, \mathbf{y}_j, \boldsymbol{\theta}^{(t)}] &= \int \cdots \int w_{cj} I_A(\mathbf{w}_j) \left(\nu_{1j}^{(t)}\right)^{w_{1j}} \left(1-\nu_{1j}^{(t)}\right)^{1-w_{1j}} d\mathbf{w}_j \\
&= \nu_{1j}^{(t)}.
\end{aligned}
$$

□

Theorem B.2.3. *The posterior density of $\boldsymbol{\gamma}_j$ is*

$$f(\boldsymbol{\gamma}_j|\mathbf{y}_j,\boldsymbol{\theta}^{(t)}) = \big(\eta_j^{(t)}\big)^{\gamma_j}\big(1-\eta_j^{(t)}\big)^{1-\gamma_j}, \tag{B.7}$$

where we recall the definition of the posterior expectation of γ_j

$$\eta_j^{(t)} = E\big[\gamma_j|\mathbf{y}_j,\boldsymbol{\theta}^{(n)}\big] = \frac{p^{(t)}f_1^{(t)}(\mathbf{y}_j)}{p^{(t)}f_1^{(t)}(\mathbf{y}_j)+(1-p^{(t)})f_2^{(t)}(\mathbf{y}_j)},$$

with

$$f_1^{(t)}(\mathbf{y}_j) = \prod_{c\in\mathcal{C}}\Big(\pi_{1c}^{(t)}\prod_{i\in c}\phi(y_{ij}|\mu_{1i}^{(t)},\sigma_{1i}^{2(t)}) + \pi_{0c}^{(t)}\prod_{i\in c}\phi(y_{ij}|\mu_{2i}^{(t)},\sigma_{2i}^{2(t)})\Big),$$

and

$$f_2^{(t)}(\mathbf{y}_j) = \pi_1^{(t)}\prod_{i=1}^{n}\phi(y_{ij}|\alpha_{1i}^{(t)},\varsigma_{1i}^{2(t)}) + \pi_0^{(t)}\prod_{i=1}^{n}\phi(y_{ij}|\alpha_{2i}^{(t)},\varsigma_{2i}^{2(t)}).$$

Proof. From (B.2) and (B.3) we get

$$\begin{aligned}
& f(\boldsymbol{\gamma}_j|\mathbf{y}_j,\boldsymbol{\theta}^{(t)}) \\
= \;& \frac{f(\mathbf{y}_j,\boldsymbol{\gamma}_j|\boldsymbol{\theta}^{(t)})}{f(\mathbf{y}_j|\boldsymbol{\theta}^{(t)})} \\
= \;& \frac{\Big\{p^{(t)}f_1^{(t)}(\mathbf{y}_j)\Big\}^{\gamma_j}\Big\{\big(1-p^{(t)}\big)f_2^{(t)}(\mathbf{y}_j)\Big\}^{1-\gamma_j}}{p^{(t)}f_1^{(t)}(\mathbf{y}_j)+(1-p^{(t)})f_2^{(t)}(\mathbf{y}_j)} \\
= \;& \left\{\frac{p^{(t)}f_1^{(t)}(\mathbf{y}_j)}{p^{(t)}f_1^{(t)}(\mathbf{y}_j)+(1-p^{(t)})f_2^{(t)}(\mathbf{y}_j)}\right\}^{\gamma_j}\left\{\frac{\big(1-p^{(t)}\big)f_2^{(t)}(\mathbf{y}_j)}{p^{(t)}f_1^{(t)}(\mathbf{y}_j)+(1-p^{(t)})f_2^{(t)}(\mathbf{y}_j)}\right\}^{1-\gamma_j},
\end{aligned}$$

which is that of (B.7). □

We now have everything we need to formally derive the Q-function. Recall the complete data loglikelihood given in (8.8). We need to calculate $E_{\mathbf{w}_j,\gamma_j|\mathbf{y}_j}[w_{cj}\gamma_j]$, for all c, $E_{\mathbf{w}_j,\gamma_j|\mathbf{y}_j}[(1-\gamma_j)w_{1j}]$, and $E_{\mathbf{w}_j,\gamma_j|\mathbf{y}_j}[(1-\gamma_j)\log I_A(\mathbf{w}_j)]$. We get

1.

$$\begin{aligned} E\big[\gamma_j w_{cj} \big| \mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] &= E\big[\gamma_j E[w_{cj}|\gamma_j, \mathbf{y}_j, \boldsymbol{\theta}^{(t)}]\big|\mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \\ &= E\big[\gamma_j\big(\gamma_j\tau_{cj}^{(t)} + (1-\gamma_j)\nu_{1j}^{(t)}\big)|\mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \\ &= \tau_{cj}^{(t)} E\big[\gamma_j|\mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \\ &= \tau_{cj}^{(t)}\eta_j^{(t)}, \end{aligned}$$

since $\gamma_j^2 = \gamma_j$ and $\gamma_j(1-\gamma_j) = 0$.

2.

$$\begin{aligned} E\big[(1-\gamma_j) w_{cj} \big| \mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] &= E\big[(1-\gamma_j) E[w_{cj}|\gamma_j, \mathbf{y}_j, \boldsymbol{\theta}^{(t)}]\big|\mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \\ &= E\big[(1-\gamma_j)\big(\gamma_j\tau_{cj}^{(t)} + (1-\gamma_j)\nu_{1j}^{(t)}\big)|\mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \\ &= \nu_{1j}^{(t)} E\big[(1-\gamma_j)|\mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \\ &= \nu_{1j}^{(t)}(1-\eta_j^{(t)}) \end{aligned}$$

3.

$$\begin{aligned} & E\big[(1-\gamma_j)\log I_A(\mathbf{w}_j)\big|\mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \\ = \; & P\big[\gamma_j = 0|\mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \cdot E\big[(1-\gamma_j)\log I_A(\mathbf{w}_j)\big|\gamma_j = 0, \mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \\ & + P\big[\gamma_j = 1|\mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \cdot E\big[(1-\gamma_j)\log I_A(\mathbf{w}_j)\big|\gamma_j = 1, \mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \\ = \; & P\big[\gamma_j = 0|\mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \cdot E\big[\log I_A(\mathbf{w}_j)\big|\gamma_j = 0, \mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \\ = \; & P\big[\gamma_j = 0|\mathbf{y}_j, \boldsymbol{\theta}^{(t)}\big] \int I_A(\mathbf{w}_j)\big(\nu_{1j}^{(t)}\big)^{w_{1j}}\big(1-\nu_{1j}^{(t)}\big)^{1-w_{1j}}\log I_A(\mathbf{w}_j)d\mathbf{w}_j \\ = \; & 0, \end{aligned}$$

since $I_A(\mathbf{w}_j)\log I_A(\mathbf{w}_j) = 0\cdot(-\infty) = 0$ or $I_A(\mathbf{w}_j)\log I_A(\mathbf{w}_j) = 1\cdot 0 = 0$.

Now plug into (8.8) and we arrive at the Q-function in (8.11)

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{j=1}^{G}\Big(\eta_j^{(t)}\log p + (1-\eta_j^{(t)})\log(1-p)\Big) \\
&+ \sum_{j=1}^{G}\sum_{c\in\mathcal{C}}\Big(\eta_j^{(t)}\tau_{cj}^{(t)}\log\pi_{1c} + \eta_j^{(t)}(1-\tau_{cj}^{(t)})\log\pi_{0c}\Big) \\
&+ \sum_{j=1}^{G}\sum_{c\in\mathcal{C}}\Big(\eta_j^{(t)}\tau_{cj}^{(t)}\sum_{i\in c}\log\phi(y_{ij}|\mu_{1i},\sigma_{1i}^2) \\
&\qquad +\eta_j^{(t)}(1-\tau_{cj}^{(t)})\sum_{i\in c}\log\phi(y_{ij}|\mu_{2i},\sigma_{2i}^2)\Big) \\
&+ \sum_{j=1}^{G}\Big((1-\eta_j^{(t)})\nu_{1j}^{(t)}\log\pi_1 + (1-\eta_j^{(t)})(1-\nu_{1j}^{(t)})\log\pi_0\Big) \\
&+ \sum_{j=1}^{G}\Big((1-\eta_j^{(t)})\nu_{1j}^{(t)}\sum_{i=1}^{n}\log\phi(y_{ij}|\alpha_{1i},\varsigma_{1i}^2) \\
&\qquad +(1-\eta_j^{(t)})(1-\nu_{1j}^{(t)})\sum_{i=1}^{n}\log\phi(y_{ij}|\alpha_{2i},\varsigma_{2i}^2)\Big).
\end{aligned}
\tag{B.8}
$$

B.2.2 M-step

In maximizing the Q-function in (B.8) we differentiate with respect to $(p, \pi_1, (\pi_{1c})_{c\in\mathcal{C}})$ and $((\mu_{1i}, \sigma_{1i}^2, \mu_{2i}, \sigma_{2i}^2)_i, (\alpha_{1i}, \varsigma_{1i}^2, \alpha_{2i}, \varsigma_{2i}^2)_i)$, where

$$
\log\phi(y_{ij}|\mu_{ki},\sigma_{ki}^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_{ki}^2) - \frac{1}{2\sigma_{ki}^2}(y_{ij}-\mu_{ki})^2, \quad k=1,2,
$$

and

$$
\log\phi(y_{ij}|\alpha_{ki},\varsigma_{ki}^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\varsigma_{ki}^2) - \frac{1}{2\varsigma_{ki}^2}(y_{ij}-\alpha_{ki})^2, \quad k=1,2.
$$

Differentiating with respect to p and setting to zero we get

$$
\sum_{j=1}^{G}\left(\frac{\eta_j^{(t)}}{p} - \frac{1-\eta_j^{(t)}}{1-p}\right) = 0,
$$

which leads to

$$p^{(t+1)} = \frac{1}{G}\sum_{j=1}^{G}\eta_j^{(t)}.$$

Differentiating with respect to π_{1c}, for all $c \in \mathcal{C}$, and setting to zero we get

$$\sum_{j=1}^{G}\left(\frac{\eta_j^{(t)}\tau_{cj}^{(t)}}{\pi_{1c}} - \frac{\eta_j^{(t)}(1-\tau_{cj}^{(t)})}{1-\pi_{1c}}\right) = 0,$$

which leads to

$$\pi_{1c}^{(t+1)} = \frac{\sum_{j=1}^{G}\eta_j^{(t)}\tau_{cj}^{(t)}}{\sum_{j=1}^{G}\eta_j^{(t)}}.$$

Differentiating with respect to π_1 and setting to zero we get

$$\sum_{j=1}^{G}\left(\frac{(1-\eta_j^{(t)})\nu_{1j}^{(t)}}{\pi_1} - \frac{(1-\eta_j^{(t)})(1-\nu_{1j}^{(t)})}{1-\pi_1}\right) = 0,$$

which leads to

$$\pi_1^{(t+1)} = \frac{\sum_{j=1}^{G}(1-\eta_j^{(t)})\nu_{1j}^{(t)}}{\sum_{j=1}^{G}(1-\eta_j^{(t)})},$$

Differentiating with respect to μ_{1i}, for $i \in c$, and setting to zero we get

$$\sum_{j=1}^{G}\frac{\eta_j^{(t)}\tau_{cj}^{(t)}}{\sigma_{1i}^2}\big(y_{ij} - \mu_{1i}\big) = 0,$$

which leads to

$$\mu_{1i}^{(t+1)} = \frac{\sum_{j=1}^{G}\eta_j^{(t)}\tau_{cj}^{(t)}y_{ij}}{\sum_{j=1}^{G}\eta_j^{(t)}\tau_{cj}^{(t)}}.$$

Similarly if we differentiate with respect to μ_{2i}, for $i \in c$, and set to zero we get

$$\mu_{2i}^{(t+1)} = \frac{\sum_{j=1}^{G}\eta_j^{(t)}(1-\tau_{cj}^{(t)})y_{ij}}{\sum_{j=1}^{G}\eta_j^{(t)}(1-\tau_{cj}^{(t)})}.$$

Now differentiating with respect to σ_{1i}^2, for $i \in c$, and setting to zero we get

$$-\sum_{j=1}^{G} \frac{\eta_j^{(t)} \tau_{cj}^{(t)}}{2\sigma_{1i}^2} + \sum_{j=1}^{G} \frac{\eta_j^{(t)} \tau_{cj}^{(t)} (y_{ij} - \mu_{1i}^{(t+1)})^2}{2(\sigma_{1i}^2)^2} = 0,$$

which leads to

$$\sigma_{1i}^{2(t+1)} = \frac{\sum_{j=1}^{G} \eta_j^{(t)} \tau_{cj}^{(t)} \left(y_{ij} - \mu_{1i}^{(t+1)}\right)^2}{\sum_{j=1}^{G} \eta_j^{(t)} \tau_{cj}^{(t)}}.$$

Similarly, if we differentiate with respect to σ_{2i}^2, for $i \in c$, and set to zero we get

$$\sigma_{2i}^{2(t+1)} = \frac{\sum_{j=1}^{G} \eta_j^{(t)} \left(1 - \tau_{cj}^{(t)}\right) \left(y_{ij} - \mu_{2i}^{(t+1)}\right)^2}{\sum_{j=1}^{G} \eta_j^{(t)} \left(1 - \tau_{cj}^{(t)}\right)}$$

With the same procedure as above, by differentiating with respect to $\alpha_{1i}, \alpha_{2i}, \varsigma_{1i}^2, \varsigma_{2i}^2$, for all i, we get the updating formulas

$$\begin{aligned} \alpha_{1i}^{(t+1)} &= \frac{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) \nu_{1j}^{(t)} y_{ij}}{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) \nu_{1j}^{(t)}}, \\ \alpha_{2i}^{(t+1)} &= \frac{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) (1 - \nu_{1j}^{(t)}) y_{ij}}{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) (1 - \nu_{1j}^{(t)})}, \end{aligned} \tag{B.9}$$

and

$$\begin{aligned} \varsigma_{1i}^{2(t+1)} &= \frac{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) \nu_{1j}^{(t)} \left(y_{ij} - \alpha_{1i}^{(t+1)}\right)^2}{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) \nu_{1j}^{(t)}}, \\ \varsigma_{2i}^{2(t+1)} &= \frac{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) (1 - \nu_{1j}^{(t)}) \left(y_{ij} - \alpha_{2i}^{(t+1)}\right)^2}{\sum_{j=1}^{G} (1 - \eta_j^{(t)}) (1 - \nu_{1j}^{(t)})}. \end{aligned} \tag{B.10}$$

BIBLIOGRAPHY

Aalen, O. O. (1980), "A model for non-parametric regression analysis of counting processes," *Lecture Notes in Statistics-2: Mathematical Statistics and Probability Theory, (W. Klonecki, A. Kozek, and J. Rosinski, eds.), Springer Verlag: New York*, pp. 1–25.

Banfield, J. D. and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821.

Baxter, J., Jones, R., Lin, M., and Olsen, J. (2004), "SLLN for Weighted Independent Identically Distributed Random Variables," *Journal of theoretical probability*, 17, 165–181.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

Besag, J. G. (1974), "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *Journal of the Royal Statistical Society. B*, 34, 192–236.

Booth, J. G., Casella, G., and Hobert, J. P. (2008), "Clustering using objective functions and stochastic search," *Journal of the Royal Statistical Society. Series B (Methodological)*, 70, 119–139.

Chandler, R. E. and Bate, S. (2007), "Inference for clustered data using the independence log-likelihood," *Biometrika*, 94, 167–183.

Cheeseman, P. and Stutz, J. (1995), "Bayesian Classification (AutoClass): Theory and Results," *in Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press*, 49, 153–180.

Demidenko, E. (2004), *Mixed Models, Theory and Applications*, Wiley Series in Probability and Statistics.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society*, 39, 1–38.

Edwards, D. (2000), *Introduction to Graphical Modelling, 2nd edition*, Springer texts in Statistics.

Efron, B. (1988), "Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve," *Journal of the American Statistical Association*, 83, 414–425.

Figueroa, M. E., Reimers, M., Thompson, R. F., Ye, K., Li, Y., Selzer, R. R., Fridriksson, J., Paietta, E., Wiernik, P., Green, R. D., Greally, J. M., and Melnick, A. (2008), "An Integrative Genomic and Epigenomic Approach for the Study of Transcriptional Regulation," *PLoS One*, 3, e1882.

Figueroa, M. E., Wouters, B. J., and Skrabanek, L. (2009), "Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features." *Journal of the American Statistical Association*, 113, 2795–2804.

Figueroa, M. E., Lugthart, S., Li, Y., Erpelinck-Verschueren, C., Deng, X., Christos, P. J., Schifano, E., Booth, J., van Putten, W., Skrabanek, L., Campagne, F., Mazumdar, M., Greally, J. M., Valk, P. J., Lowenberg, B., Delwelsend, R., and Melnick, A. (2010), "Epigenetic signatures identify biologically distinct subtypes in acute myeloid leukemia," *Cancer Cell*, 17.

Fosen, J., Ferkingstad, E., Borgan, O., and Aalen, O. O. (2006), "Dynamic path analysis - a new approach to analyzing time-dependent covariates." *Lifetime Data Analysis*, 12, 143–167.

Fraley, C. and Raftery, A. E. (1998), "How Many clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis," *The Computer Journal*, 41, 578–588.

Fraley, C. and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

Friedman, J. H. and Meulman, J. J. (2003), "Clustering Objects on Subsets of Attributes." *technical report. Stanford University. Dept. of Statistics and Stanford Linear Accelerator Center.*

Ghosh, D. and Chinnaiyan, A. M. (2002), "Mixture Modelling of Gene Expression Data From Microarray Experiments." *Bioinformatics*, 18, 275–286.

Gueorguieva, R. V. and Agresti, A. (2001), "A correlated probit model for multivariate repeated measures of mixtures of binary and con- tinuous responses," *Journal of the American Statistical Association*, 96, 1102–1112.

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006), "A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves." *Journal of the American Statistical Association*, 101, 18–29.

Kalbfleisch, J. D. and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, Wiley Series in Probability and Statistics.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006), "Variable selection in clustering via Dirichlet process mixture models," *Biometrika*, 93, 877–893.

Lauritzen, S. L. (1996), *Graphical Models*, Clarendon, Oxford.

Lindsay, B. G. (1988), "Composite likelihood methods," *Contemporary Mathematics*, 80, 221–240.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, Wiley.

McLachlan, G. J. and Basford, K. E. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

McLachlan, G. J., Bean, R. W., and Peel, D. (2002), "A Mixture Model-Based Approach to the Clustering of Microarray Expression Data." *Bioinformatics*, 18, 413–422.

Ruppert, D., Wand, M., and Carroll, R. (2003), *Semiparametric Regression*, Cambridge University Press.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005), "Bayesian variable selection in clustering high-dimensional data," *Journal of the American Statistical Association*, 100, 602.